Behavioral flexibility is related to exploration and persistence, but not boldness or motor diversity

McCune KB[1,2] Lukas D[3] MacPherson M[1,4] Logan CJ[3]

2024-11-08

Affiliations:

1) Institute for Social, Behavioral and Economic Research, University of California Santa Barbara
2) College of Forestry, Wildlife and Environment, Auburn University
3) Max Planck Institute for Evolutionary Anthropology
4) Department of Biological Sciences, Western Illinois University

*Corresponding author: kelseybmccune@gmail.com

This is the post-study manuscript of the preregistration that was pre-study peer reviewed and received an In Principle Recommendation on 27 March 2019 by:

Jeremy Van Cleve (2019) Probing behaviors correlated with behavioral flexibility. *Peer Community in Ecology*, 100020. 10.24072/pci.ecology.100020. Reviewers: two anonymous reviewers

Preregistration: html, pdf, rmd

Post-study manuscript (submitted to PCI Ecology for post-study peer review on 11 November 2024): rmd

Abstract

Behavioral flexibility, the ability to change behavior when circumstances change based on learning from previous experience, is thought to play an important role in a species' ability to successfully adapt to new environments and expand its geographic range. However, behavioral flexibility is rarely directly tested at the individual level. This limits our ability to determine how it relates to other traits, such as exploration or persistence, that might also influence individual responses to novel circumstances. Without this information, we lack the power to predict which traits facilitate a species' ability to adapt behavior to new environments. We use great-tailed grackles (a bird species; hereafter "grackles") as a model to investigate this question because they have rapidly expanded their range into North America over the past 140 years. We evaluated whether grackle behavioral flexibility (measured as color reversal learning) correlated with individual differences in the exploration of new environments and novel objects, boldness towards known and novel threats, as well as persistence and motor diversity in accessing a novel food source. We determined that exploration of a novel environment across two time points and persistence when interacting with several different novel apparatuses was repeatable in individual grackles. There was a significant positive relationship between persistence and the two components of flexibility - the rate of learning to prefer a color option in the reversal learning task, and the rate of deviating from a preferred option. Furthermore, grackles that underwent serial reversal learning to experimentally increase behavioral flexibility were more exploratory in that they spent more time in close proximity to the novel environment relative to control individuals. This indicates that, the more an individual investigated or interacted with a novel apparatus, the more it was able to potentially learn and update its knowledge of current reward contingencies to adapt behavior accordingly. Our findings improve our understanding of the traits that are linked with flexibility in a highly adaptable species. We highlight the importance of using multiple different methods for measuring boldness and exploration to evaluate consistency of performance and therefore the methodological validity. We also show the importance of persistence as a factor in adapting to novel environmental changes.

Keywords: behavioral flexibility, personality, anthropogenic change, repeatability

Video summary https://youtu.be/Xd_nYV9Lj7E

Introduction

Humans are altering all ecosystems on the planet too rapidly for most species to evolve adaptations to survive and reproduce (Hendry et al., 2008; Sih, 2013). Among other consequences, anthropogenic change can lead to a proliferation of novel habitats, foods, and predators (Sih et al., 2011). Across short timescales, individuals must adapt to this novelty through changes in behavior. Behavioral flexibility is defined as the ability to use learning to functionally change behavior when circumstances change (Mikhalevich et al., 2017). As such, behavioral flexibility is thought to facilitate species resilience to anthropogenic change (Sol et al., 2013) and species invasions into novel areas (Sol et al., 2002; Wright et al., 2010).

Behavioral flexibility is rarely directly tested at the individual level. Research studying the impact of flexibility on the success of species invasions most often uses proxies of flexibility such as species brain size, or presence of the theoretical outcomes of flexible behavior like the number of foraging innovations (Sol et al., 2002). Until recently, few studies had directly tested the relationship between flexibility, foraging behavior, and other cognitive traits like innovativeness (Chow et al., 2016; Logan, 2016a; Audet et al., 2024). New evidence suggests that the more flexible great-tailed grackles showed greater foraging diversity in the wild (Logan et al., 2024), and were better able to innovate solutions on a novel foraging apparatus (Logan et al., 2023). Consequently, behavioral flexibility may show variation within, as well as among, species and may affect diverse aspects of individual behavioral interactions with the environment. To better understand how behavioral flexibility might facilitate responses to novelty and resilience to anthropogenic change, it is important to directly test flexibility and relate it to other ecological and behavioral traits at the individual level.

Although behavioral flexibility has been the trait that research has focused on to understand how behavioral traits 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 (Sol et al., 2002; Logan, 2016b). For example, exploration is theoretically important for increasing the likelihood of encountering fitness-enhancing resources in novel environments (Canestrelli et al., 2016; Griffin et al., 2016) and so the ability to adapt behavior to novel circumstances could be driven by exploratory tendency rather than, or in conjunction with, behavioral flexibility (Cohen et al., 2007). 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 (Carter et al., 2013).

Experimental evaluation of the relationship between flexibility and other behavioral traits has produced inconsistent results (Logan, 2016b; Dougherty & Guillette, 2018). In one well studied avian group, the Paridae, exploration is related to flexibility, implying that they are not two distinct traits, but the direction of the relationship is inconsistent across species Rojas-Ferrer et al. (2020). Inconsistencies such as this exist for other behavioral traits as well (see Logan, 2016b for detailed review). Individuals approaching a potentially threatening aspect of the environment require a certain degree of boldness (McCune et al., 2018). However, the relationship between boldness and flexibility can be positive (Titulaer et al., 2012), negative (Bebus et al., 2016; Bensky & Bell, 2022), or neutral (Guenther et al., 2014; De Meester et al., 2022). 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 (Morand-Ferron et al., 2022). 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 (Diquelou et al., 2015). Research in squirrels supports this prediction (Chow et al., 2016), 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 (Logan, 2016b).

The lack of consistent support for which behavioral traits are related (or not) to flexibility could stem from what has been called a "jingle-jangle fallacy" (Carter et al., 2013). This term describes the mismatch between a trait label (like exploration) and what the method (novel environment) actually measures (could be exploration, activity, or boldness). A mismatch can occur when researchers use a single trait label for what are actually multiple distinct inherent traits ("jingle fallacy"), or if using distinct labels for what is actually the same inherent trait ("jangle fallacy"). One step towards avoiding this issue is to use multiple experimental methods, as in a test battery, to measure a variety of behaviors, then assess the relationships among performance to identify which aspects of the behaviors that are measured might be driven by the same underlying trait (Perals et al., 2017; Shaw & Schmelz, 2017).

To determine whether behavior labels represent the same underlying trait, it is also important to ensure that measured performance on behavioral assays is consistent within individuals across time and context (i.e., repeatable). Inter-individual differences in performance could result from short-term variation in the external environment like social interactions or food availability. Furthermore, short-term differences in internal states like hunger or stress can lead to variation in behavior within species. This plasticity is distinct from consistent individual differences in behavior across contexts stemming from genetic or developmental effects (i.e., animal personality; Duckworth (2010); Fidler et al. (2007)). Only behaviors that stem from inherent characteristics can be evolutionarily linked through natural selection (Réale et al., 2007; Rowe & Healy, 2014). It is important to know whether traits are linked because such linkage could result in limited behavioral plasticity that may alter the ability or mode of adapting to rapid environmental changes (Sih et al., 2004). Indeed, inconsistency in the direction of the relationship between flexibility and behavioral traits in previous studies could stem from a lack of repeatability in performance on behavioral trait assays. To address whether behavioral flexibility is related to other behavioral traits, we must first assess whether our methods produce performance that is repeatable (Dingemanse & Dochtermann, 2013) to validate that it is more likely to represent variation in an inherent trait.

Here, we first test whether performance on measures of exploration, boldness, persistence, and motor diversity is repeatable across time and contexts and therefore likely represents distinct personality traits. Behavior is considered repeatable if the variance in performance on the task is smaller within individuals compared to the variance among individuals. If there is no repeatability of these behaviors within individuals, then performance is likely state dependent (i.e., it depends on fluctuating motivation, stress, hunger levels, etc.) and/or reliant on the current context of the tasks. Then we assessed whether the repeatable traits are related to performance on a behavioral flexibility task. We focus on great-tailed grackles (*Quiscalus mexicanus*; hereafter "grackles") because they are likely to have experienced selection for behavioral adaptations to rapid environmental change. Grackles have rapidly expanded their range into novel areas in North America over the past 140 years Summers et al. (2023) and our previous research on this species has demonstrated that grackles are behaviorally flexible (Logan, 2016c), and that behavioral flexibility is a distinct trait on which grackles show individual variation (McCune et al., 2023). Thus, this species is ideal for assessing whether behavioral flexibility is part of a suite of behaviors that facilitate adaptation to novel environments.

Preregistered hypotheses and predictions summary

We preregistered several additional predictions pertaining to alternative measures of behavioral flexibility that we are not using here. The preregistration details the criteria that determined which variables to use, and we summarize this below in Methods > Behavioral flexibility. The full preregistration is available *as Supplementary Material 3. The prediction numbers listed here maintain the original order from the preregistration to help readers track consistency across Stage 1 and Stage 2.*

Hypothesis 1: Behavioral flexibility is correlated with the exploration of new environments and novel objects, but not with boldness, persistence, or motor diversity.

Predictions 1-5: Behaviorally flexible individuals will be more exploratory of novel environments (P1) and novel objects (P2) than less flexible individuals, but there will be no difference in persistence (P3), boldness (P4), or motor diversity (P5) (as found in Logan, 2016b).

P1 alternative 4: There is no correlation between exploration and behavioral flexibility because our novel object and novel environment methods are inappropriate for measuring exploratory tendency. These measures of exploration both incorporate novelty and thus may measure boldness rather than exploration. This will be supported by a positive correlation between behavioral responses to our exploration and boldness assays.

P3 alternative 1: There is a positive correlation between persistence and the number of incorrect choices in reversal learning before making the first correct choice. This indicates that individuals that are persistent in one context are also persistent in another context.

P3 alternative 2: There is no correlation between persistence and the number of incorrect choices in reversal learning before making the first correct choice. This indicates that flexibility is an independent trait.

Hypothesis 2: Captive and wild individuals may respond differently to assays measuring exploration and boldness.

Prediction 6: Individuals assayed while in captivity are less exploratory and bold than when they are again assayed in the wild, and as compared to separate individuals assayed in the wild, potentially because captivity is an unfamiliar situation.

P6 alternative 1: Individuals in captivity are more exploratory and bold than wild individuals (testing sessions matched for season), and captive individuals show more exploratory and bold behaviors than when they are subsequently tested in the wild, potentially because the captive environment decreases the influence of predation, social interactions and competition.

P6 alternative 2: There is no difference in exploration and boldness between individuals in captivity and individuals in the wild (matched for season), potentially because in both contexts our data is biased by sampling only the types of individuals that were most likely to get caught in traps.

P6 alternative 3: Captive individuals, when tested again after being released, show no difference in exploratory and bold behaviors because our methods assess inherent personality traits that are consistent across the captive and wild contexts in this taxa.

Methods

Preregistration details

The hypotheses, methods, and analysis plan are described in detail in the peer-reviewed preregistration. We summarize these methods here, with any changes from the preregistration noted in the *Changes after the study began* section. The preregistration was written and submitted to Peer Community In (PCI) Ecology for peer review (Sep 2018) before collecting any data. After data collection began (and before any data analysis was conducted), we received peer reviews from PCI Ecology, revised, and resubmitted the preregistration (Feb 2019). It received an in principle recommendation in Mar 2019.

Summary of methods

Subjects

Grackles were caught in the wild in Tempe, Arizona USA. All individuals received color leg bands for individual identification and some individuals (n=19) were brought temporarily into aviaries. We gave these individuals various assays to measure behavioral flexibility, exploration, boldness, persistence and motor diversity, and then released them back to the wild. Grackles were individually housed in an aviary (each 244 cm long by 122 cm wide by 213 cm tall) for a maximum of six months where they had *ad lib* access to water. During testing (except exploration, see below) we food deprived grackles for up to four hours per day, but they had the opportunity to receive high value food items by participating in the assays. They had access to a maintenance diet at all other times. Individuals were given three to four days to habituate to the aviaries before their test battery began. For our second hypothesis, we tested as many grackles as possible in the wild that were color-banded (n=18 total, including 4 previously tested in the aviaries).

Behavioral flexibility

As part of a different investigation, we used serial reversal learning to measure and then increase grackle behavioral flexibility. Details on the methods and results from this research are published elsewhere (Logan et al., 2023; McCune et al., 2023). Briefly, we trained grackles to search in one of two color containers for food (Fig. 1a). After grackles showed a significant preference for this color (passing criterion was 17/20 trials correct), we switched the location of the food to the other color container (a "reversal", for which we used the same passing criterion). We measured baseline behavioral flexibility as the number of trials it took grackles to switch their preference in the first reversal and search primarily in the second color container. A randomized subset of grackles (n = 8) received training to experimentally increase behavioral flexibility. We switched the location of the food multiple times (serial reversals) until grackles were switching their preference quickly enough to meet our passing criterion of two consecutive reversals in 50 trials or fewer. Instead of serial reversals, control grackles (n = 11) received equal testing experience with two identically colored containers, both containing a food item.

In addition to assessing the relationship between performance on the behavioral trait assays and whether the grackle was flexibility trained (or in the control group), we preregistered that we would assess which of multiple additional continuous variables best represented flexible behavior. These variable were 1) the number of trials to reverse a preference in the last reversal the individual receives (for control individuals, the first reversal was also the last reversal); 2) The ratio of correct divided by incorrect trials for the first 40 trials in their final reversal, after the individual has seen the newly rewarded option once; 3) the average number of trials to solve and the average number of trials to attempt a new option on the multiaccess box (MAB), if these variables are uncorrelated with reversal performance; and 4) the "Flexibility Comprehensive" variables, described below. We determined that the Flexibility Comprehensive variables more effectively represent flexibility (Lukas et al., 2022). Furthermore, we found the latency to switch to a new locus on the MAB tasks was correlated with reversal performance. As a result, we focus only on the Flexibility Comprehensive variables here.

From the performance of each individual on reversal learning, we created 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 (Blaisdell et al., 2021; Lukas et al., 2022). This measure of flexibility includes 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.

Measures of exploration, boldness, motor diversity, and persistence were collected after the serial reversal learning experiment was complete. By experimentally increasing behavioral flexibility we increased our ability to detect a relationship between this trait and the other traits under investigation in this study.

Boldness

We define boldness as an individual's response to a potential threat (Réale et al., 2007). We measured boldness with two different threatening objects, a known threat (taxidermied Cooper's hawk) and a novel threat (purple cat halloween decoration). We also included a known non-threat (taxidermied pigeon) as a control condition (Fig. 1d). Each individual was assayed with all three objects, presented in randomized order, across three days. Exposure to each object was limited to 15 minute trials, and a food item was placed next to the object. Boldness assays occurred while the grackle was food deprived to elicit approach behaviors. We conducted each of these assays twice to measure the repeatability of performance on this task to verify that the experimental designs elicited behaviors indicative of an inherent personality trait (as opposed to a passing motivational state). During boldness trials we measured multiple behaviors and, as preregistered, statistically analyzed the variable for which we ultimately had the most data (see below).

Exploration

We defined exploration as an individual's response to novelty (Réale et al., 2007) to gather information that does not satisfy immediate needs (Mettke-Hofmann et al., 2002). We used two different assays to measure exploratory tendency: novel environment (a small tent) and novel object (a pink fuzzy shape) exploration (Fig. 1b & 1c). We also conducted control conditions where we measured the grackle's behavior in its familiar environment (the aviary) and with a familiar object (an empty water dish). Exploration tests occurred when the grackle was not food deprived to ensure that any approach to the novel object was for information gathering rather than food. Each trial was 45 minutes long and we always conducted the familiar condition trial immediately before the novel condition trial. We also conducted each of these assays twice to measure repeatability. As in boldness trials, we measured multiple behaviors during exploration trials and statistically analyzed the variable for which we ultimately had the most data (see below).

Motor diversity and persistence

We defined motor diversity as the number of different motor actions used to solve novel problems on either of two multiaccess boxes (MABs; Fig. 1e & 1f). We used an ethogram (Table 1) to define and distinguish each interaction with the MABs. We quantified persistence as the number of touches to a novel apparatus per trial time (Griffin & Diquelou, 2015; Logan, 2016b), 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. Touches to the MABs were separated 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). Motor diversity and persistence were coded from videos of grackles interacting with the two different MAB apparatuses for a separate experiment on problem solving ability (Logan et al., 2023), as well as the novel apparatuses from the exploration and boldness assays.

Individual differences assays in the wild

We attempted to measure boldness and exploration in free-flying color-banded individuals in their home ranges. The overall methods for the assays in the wild were similar to those conducted in the aviaries. However, to attract the grackles' attention to the items, we always used food near the site where the items were placed. For the exploration assays this food was greater than 2m from the item, whereas in the boldness assays the food was right next to the item (as in the aviary assays). We began the trial when a color-banded grackle came within view (20m) and was able to see the food and the object. In contrast to the aviary assays, we allowed multiple grackles to engage in the wild assays at one time. We then measured the same variables as those in the assays conducted in the aviaries. We also attempted to obtain repeated measures in the wild assays, but it was much more difficult to obtain repeated participation with free-flying grackles in the wild.

Statistical analyses

General analysis plan - For all analyses, we used the MCMCglmm function in the MCMCglmm R package (Hadfield, 2010). Our preregistered analysis plan was to use a Poisson distribution and log link for both the repeatability analyses and analyses testing the correlation of behavioral traits with flexibility. However, we used the DHARMa package (Hartig, 2019) to verify that the data for each analysis met the assumptions for Poisson regression and modified the model family accordingly (see below in "Changes after the study began"). We started each model with 13,000 iterations, a thinning interval of 10, a burnin of 3,000, and minimal priors (V=1, nu=0). We checked that the GLMM showed acceptable convergence [i.e., lag time autocorrelation values <0.01; Hadfield (2010)], and adjusted the number of iterations, thinning and burnin if necessary.

We quantified multiple variables describing performance in each of the boldness and exploration assays. However, some individuals do not show all behaviors (e.g., not all individuals enter into the novel environment). Therefore, our preregistered analysis plan states that for each assay we will choose the variable for which we have the most data to proceed with the analysis. For boldness, we had the most data for the variable "Duration on the Ground", which included both the duration spent near (within 20cm) and far (between 20cm and 100cm) from the object. For exploration of the novel environment, we had the most data for "Duration near (within 20cm)" and "Latency to first land on the ground" within 100cm of the object, so we conducted one model for each variable. For exploration of the novel object, we had the most data for "Latency to first land on the ground".

Repeatability - We obtained repeatability estimates that account for the observed and latent scales. The repeatability estimate indicates how much of the total variance, after accounting for fixed and random effects, is explained by individual differences. From the posterior distribution of the MCMCglmm model for each behavioral trait, we extracted the Bird ID random effect variance to calculate the ratio of variance accounted for by individual differences relative to total variance. We used the mean value of this ratio across all iterations for a given behavioral trait as our measure of repeatability. We used the HPDinterval function from the coda package (Plummer et al., 2020) to calculate credible intervals around our repeatability estimate.

Relationship with flexibility - If performance was repeatable across two time points in the behavioral trait assays, we investigated whether performance was correlated with either of the Flexibility Comprehensive variables (ϕ and λ). Furthermore, we analyzed 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. We preregistered that we would include in these models an independent variable accounting for age effects if our subjects include juveniles as well as adults (but see *Changes after study began* section).

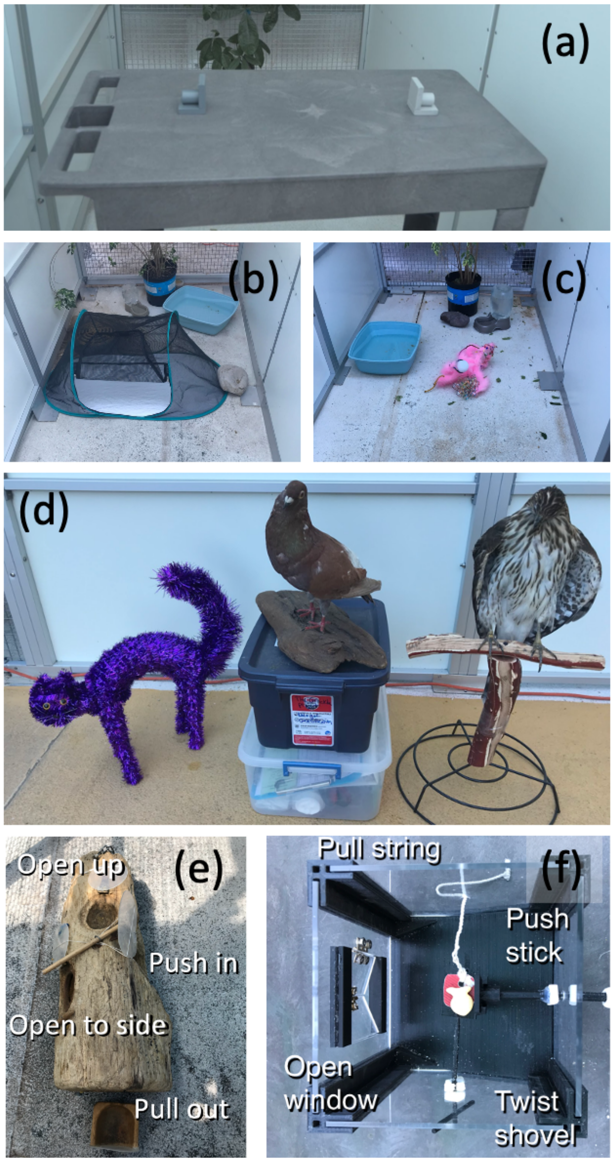
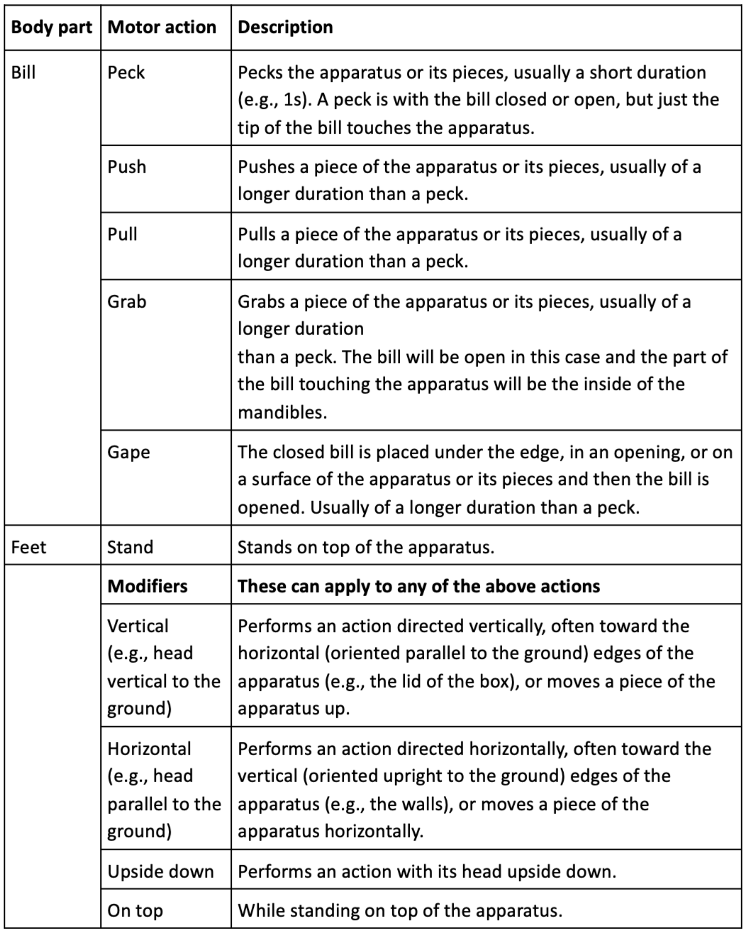


Figure 1: This experiment assessed the relationship between multiple different behavioral tests and contexts. We quantified and increased behavioral flexibility with serial reversal learning of a color preference: a light gray and a dark gray tube (a), we determined individual differences in exploration of a novel environment: a tent (b), exploration of a novel object: a homemade pink fuzzy shape (c), boldness towards threatening objects (purple halloween cat and Cooper's hawk) compared to a known non-threat (pigeon) (d), we cataloged motor diversity when interacting with novel foraging problems on the two multiaccess boxes (e-f), and we measured persistence by we counting the number of touches to all novel apparatuses (b, c, d, e, and f).

Table 1. Motor action ethogram for the two multiaccess box experiments. Any of the four modifiers can be added to any of the six motor actions. However, Stand only goes with the On top modifier, resulting in a total of 21 unique motor actions. For example, Vertical Peck is a peck to a vertical surface, and Gape Upside Down is a gape with the head being held upside down. Note that one interaction can be coded in multiple categories (e.g., if a bird pulls the string first horizontally and then vertically).



Body part	Motor action	Description
Bill	Peck	Pecks the apparatus or its pieces, usually a short duration (e.g., 1s). A peck is with the bill closed or open, but just the tip of the bill touches the apparatus.
	Push	Pushes a piece of the apparatus or its pieces, usually of a longer duration than a peck.
	Pull	Pulls a piece of the apparatus or its pieces, usually of a longer duration than a peck.
	Grab	Grabs a piece of the apparatus or its pieces, usually of a longer duration than a peck. The bill will be open in this case and the part of the bill touching the apparatus will be the inside of the mandibles.
	Gape	The closed bill is placed under the edge, in an opening, or on a surface of the apparatus or its pieces and then the bill is opened. Usually of a longer duration than a peck.
Feet	Stand	Stands on top of the apparatus.
	Modifiers	**These can apply to any of the above actions**
	Vertical (e.g., head vertical to the ground)	Performs an action directed vertically, often toward the horizontal (oriented parallel to the ground) edges of the apparatus (e.g., the lid of the box), or moves a piece of the apparatus up.
	Horizontal (e.g., head parallel to the ground)	Performs an action directed horizontally, often toward the vertical (oriented upright to the ground) edges of the apparatus (e.g., the walls), or moves a piece of the apparatus horizontally.
	Upside down	Performs an action with its head upside down.
	On top	While standing on top of the apparatus.

Changes after the study began

After data collection began and before data analysis:

1) We added an *unregistered analysis* to assess interobserver reliability for the response variables to determine how repeatable our data collection was by having the videos coded by multiple coders. This unregistered analysis is described, and results reported, in the Supplementary Material 1.

After data collection and during data analysis:

1) We conducted an *unregistered analysis* to compare the grackles' responses to the familiar item with responses to the novel/threatening items in the exploration and boldness assays. The definition for boldness relates to the behavioral response to threat, so we would expect a decrease in interactions with the novel/threatening items relative to the control item. To test that this occurred, and the grackles perceived the items as threatening, we used MCMCglmm to model the effect of condition (novel or familiar item trial) on the latency to approach and the duration spent in proximity to the items in the exploration assays. We used a gaussian distribution for latency to approach and Poisson distribution for the duration spent in proximity. We included a covariate that identified whether the bird was in the serial reversal manipulation (or not) and a random effect for bird ID. The boldness data were overdispersed and zero-inflated so we used a zero-inflated negative binomial mixed model with the R package NBZIMM (Zhang & Yi, 2020). In this model, we also included a covariate for the serial reversal manipulation and a random effect for bird ID.

2) For the repeatability analyses, we preregistered that we would calculate repeatability from the ratio of variance components extracted from MCMCglmm models. We also obtained credible intervals from the posterior distribution of these models. However, repeatability is a ratio so values can never be less than zero. As such, we are not able to ascertain the significance of our repeatability values by determining whether the credible interval overlaps with zero. We conducted an *unregistered analysis* to obtain p-values indicating whether performance was significantly more repeatable than random by utilizing the built in permutation tests in the rptR package (Stoffel et al., 2017). This also ensured that repeatability values and credible intervals were consistent with the preregistered MCMCglmm methods to validate that our non-informative priors were appropriate.

3) The boldness data were zero-inflated (69% of the data were zeros) and overdispersed, such that the appropriate model for this kind of count data is a zero-inflated negative binomial model. As stated above, we used this model type in the *unregistered analysis* to compare the responses between the threatening and non-threatening contexts. To assess repeatability of performance on the boldness assays, we preregistered that we would use a MCMCglmm model with a Poisson distribution. The boldness data were not appropriate for Poisson and we do not know of a method for obtaining the variance components for the repeatability calculation from a zero-inflated negative binomial model. Consequently, for the repeatability analysis we used a logistic regression, where the response was 0 (the grackle never approached the object during boldness trials) or 1 (the grackle approached the object during boldness trials).

4) For repeatability analyses of the exploration and persistence data, we originally planned to conduct a model with a Poisson distribution. However, the data checking process detected significant zero-inflation and heteroscedasticity in the Poisson models. We log-transformed the latency to approach (for exploration) and number of touches (for persistence) for the gaussian model, which was normally distributed and not heteroscedastic, therefore we used a gaussian distribution instead.

5) When we originally submitted this preregistration, we anticipated measuring motor diversity on only one multiaccess box (MAB). However, as part of a different experiment within our overall project, we added a second, but distinct MAB. Consequently, we did not preregister a repeatability analysis for motor diversity because there would have been only one measure per bird. We ultimately collected data on the number of motor actions on both MABs for 14 of the 17 grackles in our MAB sample, and so here we added an *unregistered analysis* to assess motor diversity repeatability. We used a Poisson regression and included a covariate for whether the grackle was flexibility trained or not. We also included an offset for the total trial time with the MABs to control for variation in the opportunity to express motor behaviors.

6) During the exploration environment assays, very few grackles stepped inside the tent ($n = 4$), so we did not have enough data to use the following preregistered variables in the analysis relating exploration and behavioral flexibility: Latency to enter a novel environment inside a familiar environment, Time spent in each of the different sections inside a novel environment or the corresponding areas on the floor when the novel environment is not present (familiar environment) as an interaction with the Environment Condition: activity in novel environment vs. activity in familiar environment, Time spent per section of a novel environment or in the corresponding areas on the floor when the novel environment is not present (familiar environment) as an interaction with the Environment Condition: time spent in novel environment vs. time spent in familiar environment.

7) We also realized that, because we experimentally increased reversal learning speed through serial reversal learning (Logan et al., 2023), behavioral flexibility should be the independent rather than dependent variable.

8) We found (Blaisdell et al., 2021; Lukas et al., 2022) that the "Flexibility Comprehensive" variables were much more effective at representing flexibility than the other variables we preregistered (e.g., Trials to reverse in the last reversal). Therefore, as preregistered, we only use this variable here, as described above in the methods. Because the individual's serial reversal learning treatment condition (control or trained) is accounted for in the flexibility comprehensive variable, we did not include condition as an

additional independent variable in these models. Note that we still conducted the preregistered analyses testing correlations between performance on the behavioral trait assays and whether the individual was in the control or flexibility trained group.

9) We preregistered that we would include "Age" as a covariate in our models relating performance on the behavioral trait assays to behavioral 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. However, we found that the performance of the two juveniles was within the range of performance of the adults. Therefore, to maintain greater statistical power, we decided to not include Age as a covariate.

10) We made two modifications to the analysis testing the relationship between persistence and flexibility. 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 tasks, we took the average of the number of functional touches and nonfunctional touches for each individual to use as the dependent variables in our models. Consequently, there was no potential for within-individual clustering in the data and we did not include the random effect for individual ID. Secondly, we were interested in the number of touches to novel objects per time. As such, we used a Poisson model as preregistered, but with an added offset term for trial time.

RESULTS

Repeatability

Our first goal was to assess the repeatability of grackle boldness, exploration, persistence and motor diversity behaviors across time and different contexts. We collected boldness and exploration data on 19 individuals, but 2 of these individuals did not participate in the MAB tasks and so our sample size was 17 for the repeatability of persistence and motor diversity.

Boldness

We first conducted an *unregistered analysis* to evaluate whether grackles perceived the objects presented to them during boldness trials as threatening. Relative to the pigeon control condition (the known non-threat), we found that grackles spent 55% less time on the ground within 100cm of the cat ($p = 0.00$) and 61% less time on the ground in the presence of the hawk ($p = 0.00$). There was a nonsignificant 9.5% decrease in duration on the ground in the hawk condition relative to the cat condition ($p = 0.71$). Consequently, there is evidence that the grackles perceived the cat and hawk as more threatening than the pigeon, and we only use data from the cat and hawk assays in all subsequent analyses including boldness. Despite the perceived threat, 12 out of 19 grackles spent time on the ground in the presence of the hawk and 7 out of 19 grackles spent time on the ground with the cat at some point during the 15-minute boldness trials.

Next we assessed whether grackles reacted consistently towards each threatening object across two time periods (temporal repeatability). Because the repeatability analysis was not possible with a zero-inflated negative binomial model, we instead used a binomial model where our dependent variable represented whether the duration grackles spent within 100cm of the threatening object was greater than 0 seconds (1) or not (0). We found no evidence for repeatability of performance in either the cat ($\mathit{Repeatability} = 0.18$, CI = 0.00-0.96, $p = 0.22$) or hawk ($R = 0.00$, CI = 0.00-0.44, $p = 0.48$) assays (Fig. 2). Similarly, when we considered grackle performance across the two different threatening contexts (contextual repeatability) there was also no consistency in behavioral response ($R = 0.04$, CI = 0.00-0.28, $p = 0.22$).

It is possible that habituation to the potentially threatening object occurs after the first exposure, such that individuals do not perform consistently across subsequent trials (Greggor et al., 2015; Takola et al., 2021).

To check whether this explains the lack of contextual repeatability in this behavioral trait, we conducted an *unregistered analysis* evaluating repeatability of performance in only the first trial in response to the potentially threatening contexts: cat, hawk, and novel object (see below). We still found no evidence that response to the potentially threatening objects was repeatable across these contexts (R = 0.00, CI = 0.00-0.17, p = 1; Fig. S3).

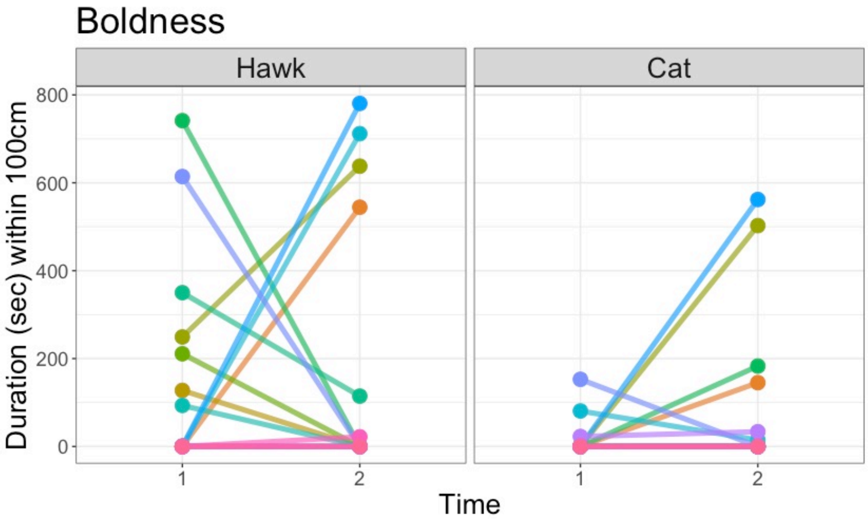


Figure 2: The grackles did not respond consistently to the threatening objects across the two time points. Each line color represents an individual and the dots show the number of seconds individuals spent on the ground within 100cm of the threatening object during each of the two 15-minute trials (Time 1 and Time 2). The two time points were separated by an average of 33 days (range: 11-49 days) and if performance was repeatable we would expect the line connecting the two dots to be at or close to horizontal.

Exploration

Similar to boldness, we assessed the repeatability of exploratory behavior across two time points and across two different contexts: a novel object and a novel environment. Because novel items might elicit a response based on the boldness personality trait rather than an exploratory response (Carter et al., 2013), we also compared the novel environment and novel object responses to control conditions with a familiar environment and a familiar object to determine whether grackles perceived the novelty as threatening (this is an *unregistered analysis*). We found no difference in the latency of individuals to approach the novel compared to the familiar environment ($ß$ = 0.29, CI = -0.24-0.81, p = 0.27), or the duration they spent near the novel and familiar environments ($ß$ = -0.61, CI = -1.47-0.20, p = 0.14). In contrast, grackles took significantly longer to approach the novel object relative to the familiar object ($ß$ = 2.11, CI = 1.22-2.89, $p < 0.01$), indicating the novel object may have been perceived as threatening.

We found that the latency to approach the novel environment across time points 1 and 2 was highly repeatable (R = 0.72, CI = 0.42-0.88, $p < 0.01$). Similarly, the duration spent near the novel environment was also highly repeatable (R = 0.85, CI = 0.67-0.98, $p < 0.01$). However, the latency to approach the novel object was not repeatable (R = 0.05, CI = 0-0.5, p = 1; Fig. 3). When we assessed performance across the novel environment and novel object tasks, we found that latency to approach was repeatable across the two different contexts, but this result was driven by the very high between-individual variance in the environment assay (R = 0.49, CI = 0.21-0.69, p = 0; Fig. S1).

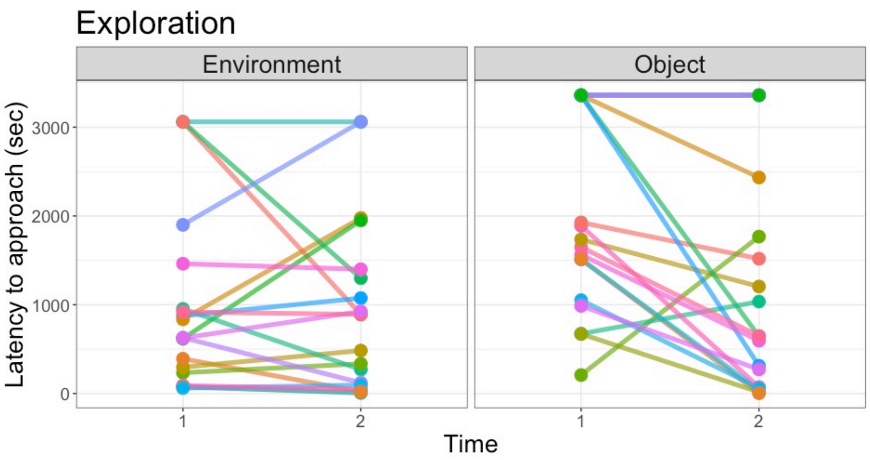


Figure 3: The response to the two exploration tests. Performance in the novel environment test was significantly repeatable across time. Grackles did not respond consistently to the novel object across the two time points. Each line color represents an individual and the dots show the amount of time before individuals approached to within 100cm of the novel item during each of the two 45-minute trials. The two time points were separated by 34 days on average (range: 11-49) and if performance is repeatable within a test we would expect the line connecting the two dots to be at or close to horizontal. Cross-contextual repeatability is indicated by similarly colored dots occurring at similar points on the y-axis across test types.

Persistence

We tested whether individuals (n = 18) were repeatable in the number of touches per trial time that they made across multiple novel test apparatuses (Fig. 1b-f): boldness objects, exploration environment and object, as well as the two different MABs. We found that persistence in interacting with these diverse objects was repeatable ($R = 0.28$, CI = 0.07-0.46, $p < 0.01$; Fig. 4)

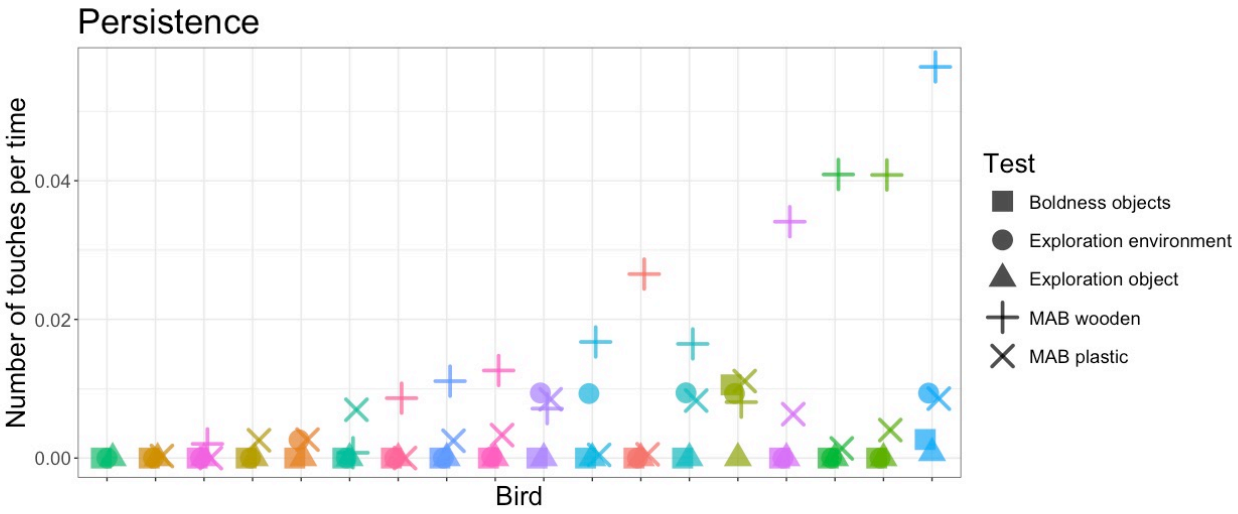


Figure 4: Persistence (the number of touches per time) was repeatable across multiple diverse test apparatuses. The x-axis shows each individual bird, which are also identified by unique colors, while the y-axis is the number of touches per trial time for each type of apparatus. Test apparatuses are distinguished by shape and we abbreviated multiaccess box as “MAB” in the figure legend.

Motor Diversity

We conducted an *unregistered analysis* to quantify repeatability in the number of different motor behaviors used while interacting with two distinct MABs in 17 grackles. Grackles were not consistent in the number of motor behaviors used across the two MABs and so repeatability was very low and not statistically significant (R = 0.06, CI = 0.00-0.45, p = 0.50).

Hypothesis 1: Relationships among measures

The repeatability analyses informed which of our methods measured consistent individual differences in behavior. Our next goal was to investigate the relationships among only the repeatable measures (exploration of a novel environment and persistence) and behavioral flexibility.

Relationship between flexibility and exploration

We first analyzed the relationship between our measures of behavioral flexibility (the Flexibility Comprehensive measure that quantifies the rate of learning to be attracted to a color option in the serial reversal learning task, ϕ, and the rate of deviating from learned associations, λ; Lukas et al. (2022); Blaisdell et al. (2021)) and two variables describing novel environment exploration: Duration near (within 20cm) the outside of the tent, and the latency to first come to the ground from the aviary perches to approach the tent. We found no relationship between either measure of novel environment exploration and ϕ or λ (Table 2).

We next investigated if performance varied as a function of whether individuals went through serial reversal learning to increase flexibility (trained group, n=8) or not (control group, n=11). Grackles that underwent the flexibility training were more exploratory in that they spent more time within 20cm of the outside of the novel environment relative to control individuals (Table 3; *ß* = 3.92, p = 0.04). However, there was no difference between trained and control individuals in latency to come to the ground within 100cm of the novel environment (*ß* = -0.43, p = 0.54).

Relationship between flexibility and persistence

In contrast, we found that grackles with a higher ϕ were more persistent in making nonfunctional touches to the novel MABs (n = 17, *ß* = 0.26, $p < 0.01$) but not functional touches (n=19, *ß* = 0.42, p = 0.11). Furthermore, individuals with lower λ were more persistent with nonfunctional (*ß* = -0.11, p =0.02) but not functional (*ß* = 0.08, p =0.77) touches. We then looked at whether the number of incorrect choices in the reversal learning task (i.e., how much the grackle is perseverating on a previously rewarded color option before exploring the other option, which is considered a measure of persistence) was related to the average number of functional or nonfunctional touches per time to the novel apparatuses (see P3 alternative 2, above). We found no evidence of a relationship between these two potential measures of persistence because the intercept-only model was supported over the model containing the number of touches variable (Table S1). This is further evidence that the number of touches is not related to perseverating on an option in a way that inhibits learning.

Also in contrast to the exploration results, we found no evidence of a relationship between persistence and whether or not the grackle underwent the flexibility training. The number of functional (*ß* = 0.81, p = 0.09) and nonfunctional touches (*ß* = 0.24, p = 0.58) to the novel apparatuses did not differ between control and trained grackles (Table 3).

Table 2: Behavioral flexibility, measured with two variables comprising our Flexibility Comprehensive measure (phi - the learning rate of attraction to either option and lambda - the rate of deviating from learned attractions), is not related to exploratory tendency as measured by duration spent within 20 cm of the outside of the novel environment (Duration Near) or the latency to approach the novel environment (Latency to Land). However, we did find that persistence (the number of

nonfunctional touches to the novel apparatuses per time) was significantly related to phi and lambda.

	Duration near			Latency to land			Number of functional touches per time			Number of nonfunctional touches per time		
	Est.	S.E.	p	Est.	S.E.	p	Est.	S.E.	p	Est.	S.E.	p
(Intercept)	2.68 [1.09, 4.26]	0.81	<0.01	6.07 [5.38, 6.77]	0.34	<0.01	-6.27 [-6.75, -5.79]	0.25	<0.01	-5.65 [-5.74, -5.57]	0.04	<0.01
c.phi[1]	1.50 [-0.18, 3.17]	0.86	0.08	-0.56 [-1.31, 0.19]	0.37	0.14	0.42 [-0.09, 0.93]	0.26	0.11	0.26 [0.17, 0.35]	0.04	**<0.01**
c.lambda[1]	-0.22 [-1.86, 1.42]	0.84	0.79	0.02 [-0.73, 0.77]	0.37	0.96	0.08 [-0.46, 0.61]	0.27	0.77	-0.11 [-0.20, -0.02]	0.05	**0.02**

[1] 'c.phi' and 'c.lambda' represent the centered and scaled version of these variables because the phi and lamda values were on fairly different scales.

Table 3: We assessed the relationship between performance on the exploration and persistence assays and whether grackles were in the serial reversal learning group where grackles were trained to be more behavioral flexibility, or in the control group. Results differed from the analysis using the Flexibility Comprehensive variables as the measure of behavioral flexibility. Grackles in the trained group were more exploratory in that they spent more time near the outside of the tent than control individuals. But no other traits were related to behavioral flexibility.

	Duration near			Latency to land			Number of functional touches per time			Number of nonfunctional touches per time		
	Est.	S.E.	p	Est.	S.E.	p	Est.	S.E.	p	Est.	S.E.	p
(Intercept)	1.16 [-0.87, 3.19]	1.04	0.26	6.26 [5.31, 7.21]	0.47	<0.01	-6.57 [-7.19, -5.95]	0.32	<0.01	-5.67 [-6.25, -5.09]	0.30	<0.01
Flexibility trained	3.61 [0.56, 6.67]	1.56	**0.02**	-0.44 [-1.90, 1.03]	0.72	0.55	0.81 [-0.12, 1.73]	0.47	0.09	0.24 [-0.61, 1.09]	0.43	0.58

Hypothesis 2: Comparing performance in captivity and in the wild

Participation of free-flying color-tagged grackles in our exploration and boldness assays in the wild was very low. Of the 19 grackles that experienced the personality assessments in the aviaries, we were only able to measure the corresponding performance in the wild for 2 in the exploration object assay and 2 in the boldness cat assay (3 individuals total). Therefore, we cannot statistically analyze the consistency of performance within individuals across aviary and wild contexts. Qualitatively, in all 4 assays in the wild, grackles approached the item more quickly in the wild compared to aviary assays (Fig. 5).

We also compared general performance on the exploration environment task (the only repeatable exploration or boldness task) of all grackles in the aviaries compared to all grackles that participated in the wild tests (i.e., many of the color-banded grackles that participated in the wild were never brought into the aviaries). We had no data from the same birds for the exploration environment test in both the aviary and wild contexts and our sample size for wild individuals was small (n=3 wild grackles, n=19 aviary grackles). From this small sample, we found no difference in the latency to approach the novel environment between the aviary or wild context (*ß* = -0.39, CI = -1.83-1.46, p = 0.63).

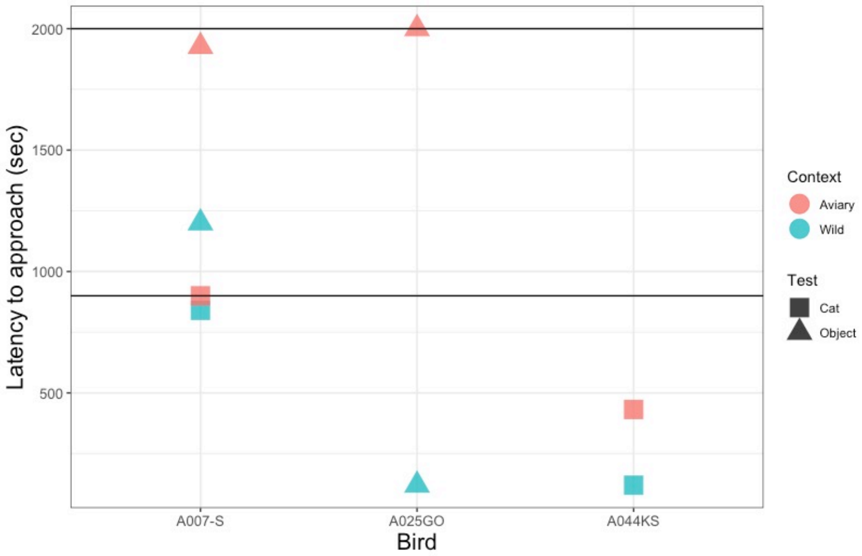


Figure 5: We were only able to measure performance on our boldness and exploration tasks on 3 individuals in both the aviaries (orange symbols) and in the wild (blue symbols). In all cases, grackles were faster to approach to within 20m of the item in the wild compared to the aviaries. The boldness cat assay is indicated with a square symbol and the exploration object assay is indicated with a triangle symbol. Note that neither of these assays produced repeatable performance across time from grackles in the aviaries. This, coupled with the small sample size means these results should be interpreted and generalized with caution. The black horizontal lines at 900 and 2000 seconds represent the ceiling values (i.e. the trial end times) for the boldness and exploration assays, respectively.

DISCUSSION

Rapid human-induced environmental change leads to novel challenges for wildlife, where individual and species ability to survive is most often possible through behavioral change (Wright et al., 2010). Although several behavioral traits are implicated in successful adaptation to human modified environments (Chapple et al., 2012), it is uncommon to directly test for multiple traits in the same individuals. Here, we used multiple novel and threatening stimuli to assess the validity of methods measuring various behavioral traits, and the relationships among traits, in great-tailed grackles, a species that has adapted to many human-induced changes to its environment during a rapid range expansion. We found that only some of our methods for measuring behavioral traits in captivity produced repeatable performance, indicating that the method elicits performance based on an inherent trait. Exploration and persistence were the two behavioral traits that were repeatable across time and context and thus are more likely to represent inherent traits that could be related to behavioral flexibility. Indeed, both exploration and persistence were correlated with behavioral flexibility.

Personality traits like boldness, exploration, and persistence are not directly observable. To validate that the experimental method used likely elicited performance reflective of the inherent personality trait, performance must be repeatable across time and contexts (Carter et al., 2013). We found that the number of touches that grackles made to multiple different novel apparatuses was repeatable, indicating that this is likely a valid method for measuring the trait persistence. Despite using multiple assays and stimuli to quantify exploration, boldness, and motor diversity, we found that only one method produced repeatable performance: the novel environment exploration assay. The other methods, exploration of a novel object, boldness towards two different novel threats, and the number of distinct motor behaviors used to interact with the two different MABs (Fig. 1) did not produce repeatable performance across sampling periods. However, we provide in Supplementary Material 2 a plot of the raw boldness and exploration data so readers can visually compare performance among tests (Fig. S2).

A key aspect distinguishing boldness from exploration is that boldness reflects a response to potentially threatening objects, novel or familiar (Carter et al., 2013; Greggor et al., 2015). Consequently, we compared performance between the novel or threatening objects and the familiar objects in the exploration and boldness assays. The novel environment was the only object the grackles did not perceive as a threat. Although the novel object for the exploration assay was not meant to be threatening (e.g., it was smaller than the threatening objects, it did not have eyes), grackles still spent significantly less time near it than their familiar object. Consequently, grackles did not perform consistently on these assays where the object was perceived as threatening. This highlights the relevance of the jingle-jangle fallacy, which describes the mismatch between a trait label and what the method actually measures (Carter et al., 2013). Although we expected the novel object to measure the trait exploration, by incorporating control conditions and multiple other novel and threatening objects, it was clear that the novel object was eliciting performance that was likely more reflective of boldness.

It is possible that grackles, in general, do not produce repeatable responses when faced with a threat in captivity. In the wild, grackles are a gregarious species that probably rarely encounters threats while alone (Johnson & Peer, 2001). For several reasons, we did not house more than one grackle in each aviary. Therefore, the lack of repeatability in performance could stem from the relatively contrived situation of experiencing a threat when visually isolated from conspecifics. While we attempted to compare performance on these personality assays between individuals in the aviaries and individuals in the wild, it was difficult to ensure participation of wild grackles. From the small sample of participating grackles, including those we also measured in the aviaries, preliminary evidence supports this explanation because wild grackles were faster to approach compared to grackles tested in the aviary. It is possible that, compared to the aviary performance, the faster approach of wild grackles could be explained by habituation to the threatening objects. If the assays in the wild occurred after grackles were released from the aviaries, it would be the third time they were exposed to the object. However, it is unlikely that this is the case because one (of three total) grackles tested in both the aviaries and the wild was actually given the novel object exploration assay and the novel threat boldness assay first in the wild, then subsequently was caught again and tested in the aviaries. This individual (A007-S) still approached the objects faster while in the wild (Fig. 5). This preliminary evidence is congruent with other research on social species encountering novelty. For example, zebra finches were more likely to approach a novel object for food (Coleman & Mellgren, 1994) and investigate a novel environment (Schuett & Dall, 2009) when in a social group compared to when alone. However, Carib grackles were slower to approach novel foraging opportunities when in a social group compared to when alone (Morand-Ferron et al., 2009). Because the majority of research on animal personality traits is conducted on individuals in captivity regardless of their sociality, more research is needed to understand when social behavior may affect the consistency of performance on personality assays.

We assessed the relationship between our repeatable behavioral traits (exploration and persistence) and the two measures of behavioral flexibility (Flexibility Comprehensive and flexibility trained versus control groups). Our Flexibility Comprehensive measure reflects two aspects of performance during serial reversal learning, the rate of learning to be attracted to a color option, ϕ, and the rate of deviating from learned associations, λ (Blaisdell et al., 2021; Lukas et al., 2022). We predicted that exploration would be positively related to flexibility, and in particular we assumed λ would best reflect exploratory behavior during the reversal learning task (Lukas et al., 2022). We found no relationship between the Flexibility Comprehensive variables and novel environment exploration. This is contrary to previous literature that found that flexibility is theoretically (Griffin et al., 2016) and experimentally (Rojas-Ferrer et al., 2020) linked with this behavioral trait. However, in support of previous literature, we found that grackles that underwent the serial reversal learning training to experimentally increase behavioral flexibility were more exploratory towards the novel environment compared to grackles that were in the control group.

We also found mixed results for the correlations between persistence and the two different measures of behavioral flexibility, though with opposite results to exploration. There was no significant correlation between persistence and whether the grackle was in the flexibility trained group. Yet, more persistent individuals had a lower λ and so were less likely to deviate from their learned attraction to the now unrewarded option. However, this relationship was only significant for nonfunctional touches. In addition, we found that persistence with nonfunctional touches was positively related to ϕ, which Lukas and colleagues (2024) determined was a better predictor of reversal performance than λ. Together, these results suggest that the more a grackle interacted with a novel apparatus, especially when no food resulted from the interaction, the more knowledge it obtained about the reward contingencies, thus facilitating faster learning of functional options. This is further supported by our finding that perseverating on choosing the previously rewarded option (incorrect choices) during reversal learning was unrelated to our persistence measure. Although persistence is often thought to impede behavioral flexibility (Morand-Ferron et al., 2022), increasingly experimental research is indicating otherwise. For example, a knockout experiment in rats found evidence for distinct cognitive mechanisms for behavioral flexibility and the inability to inhibit a response (i.e., persistence in a response; Homberg et al. (2007)). The importance of persistence in this highly adaptable species was also reported in a separate investigation by our team. Logan and colleagues (2023) found that persistence, rather than exploration or (average) flexibility, was related to the ability of a species to expand its range into novel areas because the range edge population of grackles showed higher persistence compared to the population in the center of the range (the population we focus on here).

The contradictory results for the relationship of exploration and persistence with either of the two different measures of behavioral flexibility likely reflects that individuals trained to be more flexible through serial reversal learning ended up with different strategies for how to reverse quickly (Lukas et al., 2022). Trained individuals had a higher ϕ and lower λ relative to grackles in the control group. As such, trained individuals were good at reacting to changes in the environment either because they kept on exploring alternative options (high lambda) or because they placed high importance on new information (high phi). With either strategy, we could expect trained individuals to also be better at exploration. In addition, we found that, even though all grackles improved during the training, individual differences persisted (McCune et al., 2023). These individual differences might be linked to their persistence, which would explain why the training did not influence the relationship between flexibility and persistence.

By assessing multiple behavioral traits in the same individuals of a highly adaptable species, we were able to identify correlations among certain repeatable traits that can inform our understanding of the ability to adapt to environmental change. Overall, we found that persistence, measured as the number of nonfunctional interactions with novel objects, and the time spent exploring near a novel environment are related to flexibility. Our results support previous hypotheses about traits that are related to flexible behavior, and therefore might be important for increasing survival and fitness in the face of human-induced environmental change. However, additional research is needed to further validate methods for measuring individual differences in boldness and motor diversity in this species, and to disentangle the mechanisms driving the mixed results for the relationship between persistence, exploration, and the two ways of measuring behavioral flexibility.

ETHICS

The research on the great-tailed grackles followed established ethical guidelines for the involvement and treatment of animals in experiments and received institutional approval prior to conducting the study (US Fish and Wildlife Service scientific collecting permit number MB76700A-0,1,2; US Geological Survey Bird Banding Laboratory federal bird banding permit number 23872; Arizona Game and Fish Department scientific collecting license number SP594338 [2017], SP606267 [2018], and SP639866 [2019]; Institutional Animal Care and Use Committee at Arizona State University protocol number 17-1594R; University of Cambridge ethical review process non-regulated use of animals in scientific procedures: zoo4/17 [2017]).

ACKNOWLEDGEMENTS

We thank Jeremy Van Cleve and two anonymous reviewers for their feedback on the preregistration. Julia Cissewski for tirelessly solving problems involving financial transactions and contracts; Sophie Kaube for logistical support; and Richard McElreath for project support. Ben Trumble for providing us with a wet lab at Arizona State University; Melissa Wilson Sayres for sponsoring our affiliations at Arizona State University and lending lab equipment; Kristine Johnson for technical advice on great-tailed grackles; Jay Taylor for grackle scouting at Arizona State University; Arizona State University School of Life Sciences Department

Animal Care and Technologies for providing space for our aviaries and for their excellent support of our daily activities; and our research assistants who helped habituate, trap, weigh, and train grackles to participate in tests: Nancy Rodriguez, Aelin Mayer, Sofija Savic, Brianna Thomas, Aldora Messinger, Elysia Mamola, Michael Guillen, Rita Barakat, Adrianna Boderash, Olateju Ojekunle, August Sevchick, and Justin Huynh. We thank Melissa Folsom and Luisa Bergeron for their exceptional assistance with logistical planning and data collection in the field.

SUPPLEMENTARY MATERIALS

S1 - Interobserver Reliability

Unregistered analysis: interobserver reliability of dependent variables

To determine whether the experimenter coded the dependent variables in a repeatable way, hypothesis-blind video coders were first trained in video coding the dependent variable, and then he coded 26% of the videos in the exploration and boldness experiments. We randomly chose four (Tomatillo, Queso, Mole, and Habanero) of the 19 birds (21%) who participated in these experiments using random.org. Video coders then analyzed all videos from these four birds. The experimenter's data was compared with video coder data using the intra-class correlation coefficient (ICC) to determine the degree of bias in the regression slope (Hutcheon et al. (2010), using the irr package in R: Gamer et al. (2012)).

Interobserver reliability training To pass **interobserver reliability (IOR) training**, video coders needed an ICC score of 0.90 or greater to ensure the instructions were clear and that there was a high degree of agreement across coders (see R code comments for details).

Sierra Planck (discussed with Logan): Persistence (total number of touches to apparatus) and motor diversity (presence or absence of a behavior from the ethogram). Planck was the first to code videos for these variables so there was not an already established training process or someone to compare her to. Planck and Logan worked together to agree on coding decisions using one video, and then Planck proceeded to code videos independently after that.

Alexis Breen

- **Persistence (compared with Logan):** total number of functional touches to apparatus unweighted Cohen's kappa = 1.00 (confidence boundaries=1.00-1.00, n=21 data points)
- **Persistence (compared with Logan):** total number of non-functional touches to apparatus unweighted Cohen's kappa = 0.00 (confidence boundaries=0.00-0.00, n=19 data points). Note: Breen was previously unclear about when to count non-functional touches, however, a discussion eliminated confusion and we proceeded with allowing her to video code independently because the functional touches, which she scored perfectly on, are the more difficult touches to code and thus indicative of her ability to code non-functional touches after clarity on the instructions.
- **Motor diversity (compared with Planck):** presence or absence of a behavior from the ethogram unweighted Cohen's kappa = 0.70 (confidence boundaries=0.39-1.00, n=21 data points). Note: Breen joined the project after Planck and had extensive experience with video coding bird behaviors. Because of this, and because she became Kiepsch's supervisor for exploration, boldness, persistence, and motor diversity, we decided to use Breen as the baseline for persistence and motor diversity and match future coders to her rather than to Plank. Therefore, we moved Breen into the primary video coder position (coding more of the videos than the others). To prepare for Kiepsch's training, Breen clarified the motor diversity ethogram to make it more repeatable. However, we did not require Planck to redo training because she was already so far through the videos. As such, we realize that Planck's data from 21% of the videos may not match Breen's as closely as if Plank was matched to Breen during training.

Vincent Kiepsch (compared with Breen):

- **Exploration** order of the latency-distance categories ICC = 0.96 (confidence boundaries=0.92-1.00, n=141 data points)

- **Boldness** order of the latency-distance categories ICC=1.00 (confidence boundaries=1.00-1.00, n=11 data points). Note that, for exploration and boldness, the ordered categories were aligned based on similar latencies between coders to prevent disagreements near the top of the data sheet from misaligning all subsequent entries.
  - **Persistence** number of touches to the apparatus ICC = 0.999 (confidence boundaries=0.996-1.00, n=5 data points).
- **Motor diversity**: the training score for the presence or absence of a behavior from the ethogram required additional training than originally planned, resulting in a final Cohen's kappa = 0.93 (confidence boundaries=0.80-1.00, n=42 data points).

Interobserver reliability scores were as follows (4/19 birds; 21% of the videos): Vincent Kiepsch (compared with Breen):

- **Exploration:** closest distance category to apparatus Cohen's unweighted kappa = 0.86 (confidence boundaries=0.71-1.00, n=32 data points)
- **Exploration environment:** first latency to enter tent ICC = 0.997 (confidence boundaries=0.99-0.999, n=10 data points)
- **Boldness:** closest distance to apparatus Cohen's unweighted kappa = 0.86 (confidence boundaries=0.68-1.00, n=24 data points)

Exploration and boldness in the WILD (comparison between McCune video coding and transcribing field notes for 20% of the grackles in the wild sample in March 2021 and again on the same data in May 2021): - Exploration and boldness data collected in the wild were combined because there was not much data for either and because the variables were the same for both assays - **Exploration and boldness:** closest distance category to apparatus Cohen's unweighted kappa = 1.00 (confidence boundaries=1.00-1.00, n=12 data points) - **Exploration and boldness:** latency to first landing in a distance category ICC = 0.999 (confidence boundaries=0.994-1.000, n=8 data points)

Persistence and Motor Diversity (comparisons between Breen, Kiepsch, and Planck):

- **Persistence:**
  - total number of FUNCTIONAL touches to apparatus ICC = 0.77 (confidence boundaries=0.48-0.90, n=18 data points)
  - total number of NON-FUNCTIONAL touches to apparatus ICC = 0.68 (confidence boundaries=-0.06-0.95, n=6 data points)
- **Motor diversity:** presence or absence of a behavior from the ethogram unweighted Kappa = 0.77 (confidence boundaries=0.70-0.84, n=380 data points)

These scores indicate that the dependent variables are repeatable to a moderate (persistence and motor diversity) or a high to very high (exploration and boldness) degree given our instructions and training.

S2 - Additional Boldness and Exploration Results

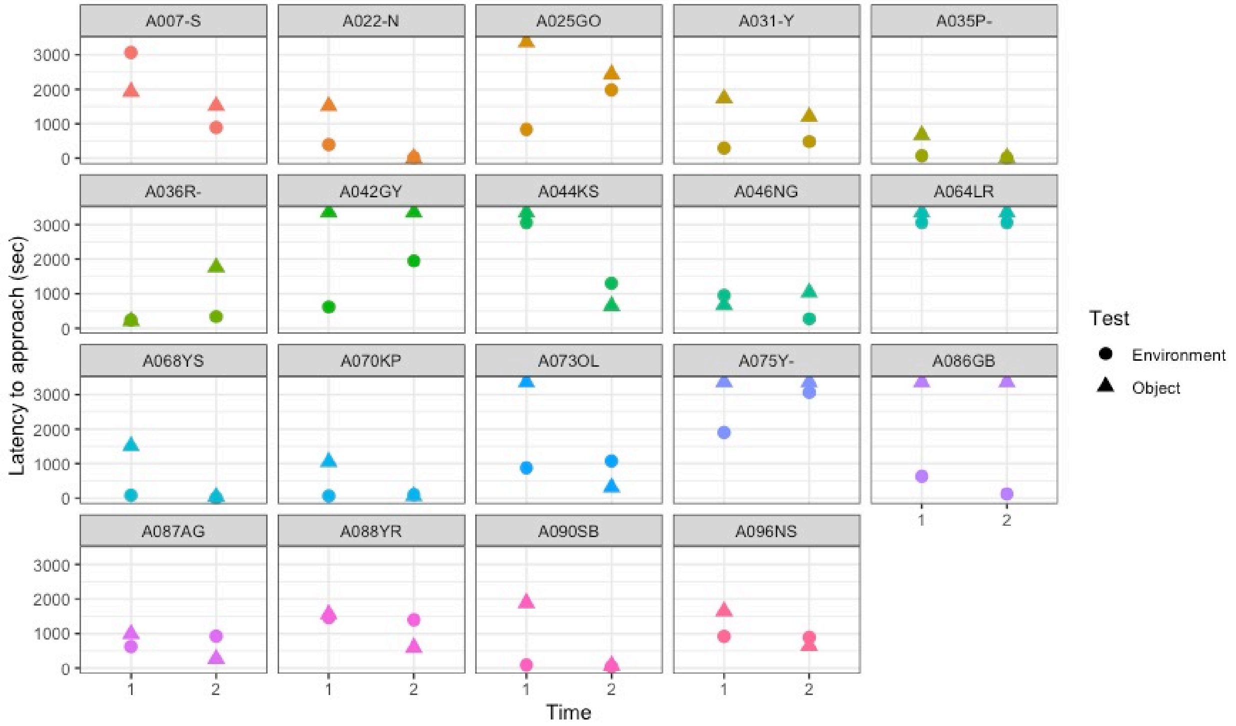


Figure S1 (repeatability of exploration): Grackles performed consistently across the two **exploration** contexts. Circles represent performance on the novel environment test and triangles represent performance on the novel object test. If performance across contexts is repeatable we would expect to see the circle and triangle at each time point to be near one another.

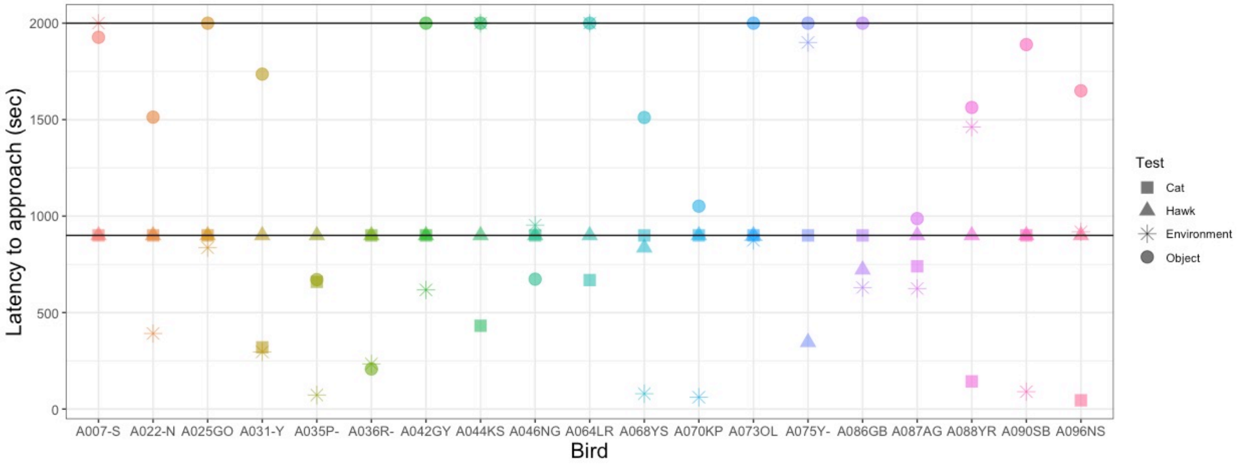


Figure S2: Performance of each grackle on the **boldness** cat (square), boldness hawk (triangle), **explore** environment (star) and explore object (circle) assays. Note that explore environment (star) was the only assay that resulted in repeatable performance across time. Here we present data only from time point one. The black horizontal lines at 900 and 2000 seconds represent the ceiling values (i.e. the trial end times) for the boldness and exploration assays, respectively.

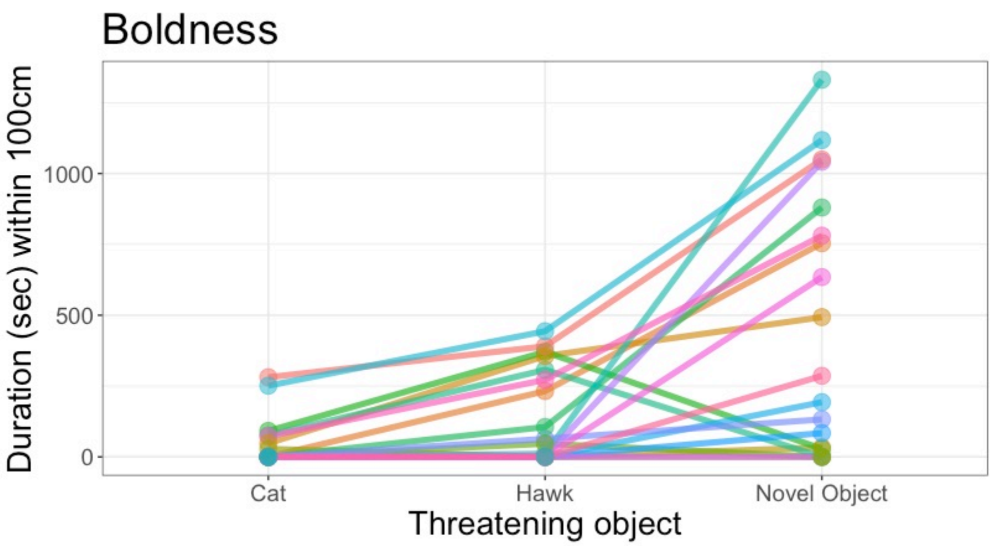
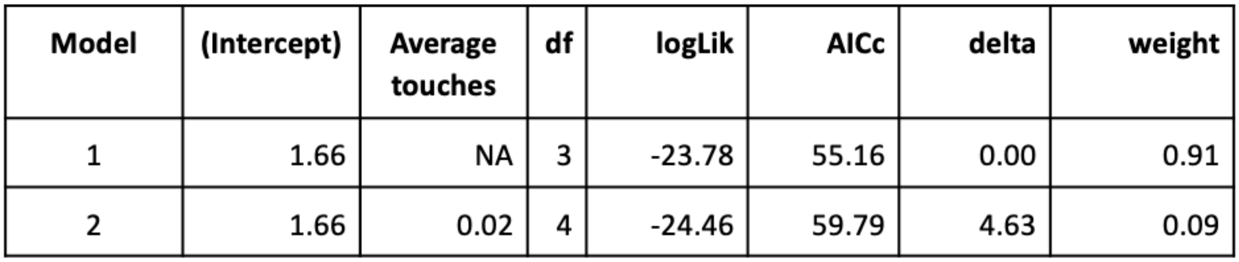


Figure S3: Habituation to the potentially threatening objects did not affect the repeatability of a grackle's response. We still found no significant repeatability in performance when only evaluating the first trial for each object. Each line color represents an individual and the dots show the number of seconds individuals spent on the ground (within 100cm) in the presence of the threatening object during Time 1's 15-minute trial.

Table S1 (hypothesis 1, prediction 3 alternative 2): Model selection output from the linear mixed model relating the number of incorrect choices on the last reversal to the average number of touches to the novel apparatuses per time. The intercept-only model (Model 1) was a better fit to the data than a model (Model 2) that included the number of touches.



Model	(Intercept)	Average touches	df	logLik	AICc	delta	weight
1	1.66	NA	3	-23.78	55.16	0.00	0.91
2	1.66	0.02	4	-24.46	59.79	4.63	0.09

S3 - Detailed Methods (Preregistration)

Below is the preregistration that passed pre-study peer review.

A. STATE OF THE DATA

NOTE: all parts of the preregistration are included in this one manuscript.

Prior to collecting any data: This preregistration was written and submitted to PCI Ecology for peer review (Sep 2018).

After data collection had begun (and before any data analysis was conducted): This preregistration was peer reviewed at PCI Ecology, revised, and resubmitted (Feb 2019), and passed pre-study peer review (Mar 2019). See the peer review history. Interobserver reliability analyses were added (Feb 2021).

B. PARTITIONING THE RESULTS

We may decide to present the results from different hypotheses in separate papers.

C. HYPOTHESES

H1: Behavioral flexibility (indicated by individuals that are faster at functionally changing their behavior when circumstances change; measured by reversal learning and switching between options on a multi-access box) is positively correlated with the exploration of new environments and novel objects, but not with other behaviors (i.e., boldness, persistence, or motor diversity) (see Mikhalevich et al. (2017) for theoretical background about our flexibility definition). We will first verify that our measures of exploration, boldness and persistence represent repeatable, inherent individual differences in behavior (i.e., personality). Individuals show consistent individual differences in behavior if the variance in latency to approach the task is smaller within individuals compared to variance in latency among individuals (for exploration and boldness assays). The same definition applies to persistence with the number of touches as the measured variable. If there is no repeatability of these behaviors within individuals, then performance is likely state dependent (e.g., it depends on their fluctuating motivation, hunger levels, etc.) and/or reliant on the current context of the tasks.

Predictions 1-5: Individuals in the experimental group where flexibility (as measured by reversal learning and on a multi-access box) was manipulated (such that individuals in the manipulated group became faster at switching) will be more exploratory of new environments (P1; methods similar to free-entry open field test as in Mettke-Hofmann et al. (2009)) and novel objects (P2; methods as in Mettke-Hofmann et al. (2009)) than individuals in the control group where flexibility was not increased, and there will be no difference between the groups in persistence (P3), boldness (P4; methods as in Logan (2016b)), or motor diversity (P5) (as found in Logan (2016b)). We do not expect the flexibility manipulation to causally change the nature of the relationship between flexibility and any of the other measured variables. Instead, we expect the manipulation to potentially enhance individual variation, thus making it easier for us to detect a correlation if one exists.

P1-P5 alternative: If the flexibility manipulation does not work in that those individuals in the experimental condition are not more flexible than control individuals, then we will analyze the individuals from both conditions as one group. In this case, we will assume that we were not able to influence their flexibility and that whatever level of flexibility they had coming into the experiment reflects the general individual variation in the population. This experiment will then elucidate whether general individual variation in flexibility relates to exploratory behaviors. The predictions are the same as above. The following alternatives apply to both cases: if the manipulation works (in which case we expect stronger effects for the manipulated group), and if the manipulation doesn't work (in which case we expect individuals to vary across all of the measured variables and for these variables to potentially interact).

P1 alternative 1: There is a positive correlation between exploration and both dependent variables in reversal learning (one accounts for exploration in reversal learning [the ratio] and the other does not). This suggests that flexibility is not independent of exploration and could indicate that another trait is present that could be explaining individual variation in flexibility as well as in exploration. This other trait or traits could be something such as boldness or persistence.

P1 alternative 2a: There is a positive correlation between exploration and the dependent variable that does not account for exploration (number of trials to reverse), but not the flexibility ratio, which suggests that performance overall in reversal learning is partially explained by variation in exploration, but that flexibility and exploration are separate traits because using a measure that accounts for exploration still shows variation in flexibility.

P1 alternative 2b: There is a negative correlation between exploration and the flexibility ratio that accounts for exploration, but not with the number of trials to reverse. This could be an artifact of accounting for exploration in both variables.

P1 alternative 3: There is no correlation between exploration and either dependent variable in reversal learning. This indicates that both dependent variables measure traits that are independent of exploration.

P1 alternative 4: There is no correlation between exploration and either dependent variable in reversal learning because our novel object and novel environment methods are inappropriate for measuring exploratory tendency. These measures of exploration both incorporate novelty and thus may measure boldness rather than exploration. This is supported by a positive correlation between behavioral responses to our exploration and boldness assays.

P3 alternative 1: There is a positive correlation between persistence and the number of incorrect choices in reversal learning before making the first correct choice. This indicates that individuals that are persistent in one context are also persistent in another context.

P3 alternative 2: There is no correlation between persistence and the number of incorrect choices in reversal learning before making the first correct choice. This indicates that flexibility is an independent trait.

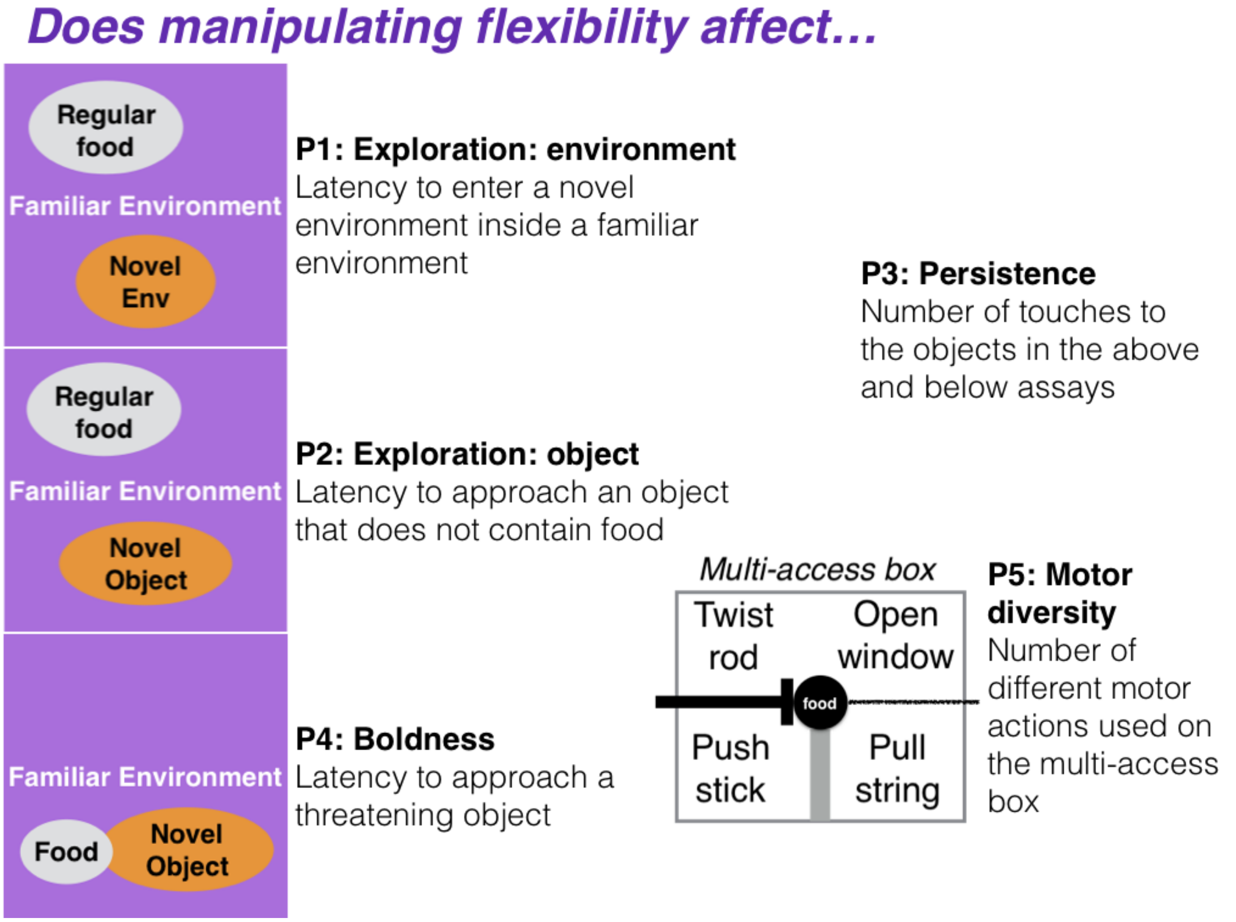


Figure 1: Figure 1.

Figure 1. An overview of the study design and a selection of the variables we will measure for each assay. Exploration will be measured by comparing individual behavior within a familiar environment to behavior towards a novel environment, as well as response to a familiar object vs. a novel object within the familiar environment that contains their regular food. Boldness will be measured as the willingness to eat next to a threatening object (familiar, novel oject, or a taxidermic predator) in their familiar environment. Persistence will be measured as the number of touches to the novel environment and novel object in the Exploration assay, the objects in the Boldness assay, and the multi-access box in a separate preregistration. Motor diversity will be measured using the multi-access box in a separate preregistration. After the flexibility manipulation occurs, assays will be conducted at least twice (e.g., Time 1, Time 2) and differences (if any)

between the control and manipulated groups in the behavioral flexibility preregistration will be compared across time and, with persistence, across tests (e.g., Test 1, Test 2) because persistence is measured in four different assays.

H2: Captive and wild individuals may respond differently to assays measuring exploration and boldness. **P6:** Individuals assayed while in captivity are less exploratory and bold than when they are again assayed in the wild, and as compared to separate individuals assayed in the wild, potentially because captivity is an unfamiliar situation.

P6 alternative 1: Individuals in captivity are more exploratory and bold than wild individuals (testing sessions matched for season), and captive individuals show more exploratory and bold behaviors than when they are subsequently tested in the wild, potentially because the captive environment decreases the influence of predation, social interactions and competition.

P6 alternative 2: There is no difference in exploration and boldness between individuals in captivity and individuals in the wild (matched for season), potentially because in both contexts our data is biased by sampling only the types of individuals that were most likely to get caught in traps.

P6 alternative 3: Captive individuals, when tested again after being released, show no difference in exploratory and bold behaviors because our methods assess inherent personality traits that are consistent across the captive and wild contexts in this taxa.

D. METHODS

Planned Sample

Great-tailed grackles are caught in the wild in Tempe, Arizona USA for individual identification (colored leg bands in unique combinations). Some individuals (~32) are brought temporarily into aviaries for testing, and then they will be released back to the wild. Grackles are individually housed in an aviary (each 244cm long by 122cm wide by 213cm tall) at Arizona State University for a maximum of three months where they have ad lib access to water at all times and are fed Mazuri Small Bird maintenance diet ad lib during non-testing hours (minimum 20h per day), and various other food items (e.g., peanuts, grapes, bread) during testing (up to 3h per day per bird). Individuals are given three to four days to habituate to the aviaries and then their test battery begins on the fourth or fifth day (birds are usually tested six days per week, therefore if their fourth day in the aviaries occurs on a day off, then they are tested on the fifth day instead). For hypothesis 2 we will attempt to test all grackles in the wild that are color-banded.

Sample size rationale

We will test as many birds as we can in the approximately three years at this field site given that the birds only participate in tests in aviaries during the non-breeding season (approximately September through March). The minimum sample size for captive subjects will be 16, however we expect to be able to test up to 32 grackles in captivity. We catch grackles with a variety of methods, some of which decrease the likelihood of a selection bias for exploratory and bold individuals because grackles cannot see the traps (i.e. mist nets). In sampling all banded birds in the wild, we will therefore have a better idea of the variation in exploration and boldness behaviors in this population.

Data collection stopping rule

We will stop testing birds once we have completed two full aviary seasons (likely in March 2020) if the sample size is above the minimum suggested boundary based on model simulations (see section “Ability to detect actual effects” below). If the minimum sample size is not met by this point, we will continue testing birds at our next field site (which we move to in the summer of 2020) until we meet the minimum sample size.

Open materials Testing protocols for exploration of new environments and objects, boldness, persistence, and motor diversity.

Open data When the study is complete, the data will be published in the Knowledge Network for Biocomplexity's data repository.

Randomization and counterbalancing There is no randomizing. The order of the three tasks will be counterbalanced across birds (using https://www.random.org to randomly assign individuals to one of three experimental orders).

1/3 of the individuals will experience:

1. Exploration environment
2. Exploration object
3. Boldness

1/3 of the individuals will experience:

1. Exploration object
2. Boldness
3. Exploration environment

1/3 of the individuals will experience:

1. Boldness
2. Exploration environment
3. Exploration object

Blinding of conditions during analysis No blinding is involved in this study. NOTE Feb 2021: inter-observer reliability analyses were conducted with hypothesis-blind video coders.

Variables included in analyses 1-5 NOTE: to view a list of these variables in a table format, please see our Google sheet, which describes whether they are a dependent variable (DV), independent variable (IV), or random effect (RE). Note: when there is more than one DV per model, all models will be run once per DV.

ANALYSIS 1 - *REPEATABILITY of boldness, persistence and exploration*

Dependent variables

1) Boldness: Latency to land on the table - OR - Latency to eat the food - OR - Latency to touch a threatening object next to food (we will choose the variable with the most data)
2) Persistence: Number of touches to an apparatus per time (multi-access box in the behavioral flexibility preregistration, novel environment in P1, and objects in P2 and P4)
3) Exploration of novel environment: Latency to enter a novel environment set inside a familiar environment
4) Exploration of novel object: Latency to land on the table next to an object (novel, familiar) (that does not contain food) in a familiar environment (that contains maintenance diet away from the object) - OR - latency to touch an object (novel, familiar) (choose the variable with the most data)

Independent variables

1) Condition: control, flexibility manipulation
2) ID (random effect because multiple measures per individual)

ANALYSIS 2 - *H1: P1-P5: flexibility correlates with exploratory behaviors*

Dependent variables

1) The **number of trials to reverse** a preference in the last reversal that individual participated in (an individual is considered to have a preference if it chose the rewarded option at least 17 out of the most recent 20 trials (with a minimum of 8 or 9 correct choices out of 10 on the two most recent sets of 10 trials)). See behavioral flexibility preregistration for details.
2) If the number of trials to reverse a preference does not positively correlate with the number of trials to attempt or solve new loci on the multi-access box (an additional measure of behavioral flexibility), then the **average number of trials to solve** and the **average number of trials to attempt** a new option on the multi-access box will be additional dependent variables. See behavioral flexibility preregistration.
3) **Flexibility comprehensive**: This measure is currently being developed and is intended be a more accurate representation of all of the choices an individual made, as well as accounting for the degree of uncertainty exhibited by individuals as preferences change. If this measure more effectively represents flexibility (determined using a modeled dataset and not the actual data), we may decide to solely rely on this measure and not use independent variables 1-3. If this ends up being the case, we will modify the code in the analysis plan below to reflect this change before conducting analyses of the data in this preregistration.

All models will be run once per dependent variable.

Independent variables

1) P1: Latency to enter a novel environment inside a familiar environment
2) P1: Time spent in each of the different sections inside a novel environment or the corresponding areas on the floor when the novel environment is not present (familiar environment) as an interaction with the Environment Condition: activity in novel environment vs. activity in familiar environment
3) P1: Time spent per section of a novel environment or in the corresponding areas on the floor when the novel environment is not present (familiar environment) as an interaction with the Environment Condition: time spent in novel environment vs. time spent in familiar environment
4) P1: Time spent exploring the outside of the novel environment (within 20cm) before entering it
5) P2: Latency to land on the table next to an object (novel, familiar) (that does not contain food) in a familiar environment (that contains maintenance diet away from the object) - OR - latency to touch an object (novel, familiar) (choose the variable with the most data)
6) P3: Number of touches to the functional part of an apparatus per time (multi-access box, novel environment in P1, novel objects in P2 and P4)
7) P3: Number of touches to the non-functional part of an apparatus per time (multi-access box)
8) P4: Latency to land on the table - OR - Latency to eat the food - OR - Latency to touch a threatening object next to food (choose the variable with the most data)
9) P5: Number of different motor actions used when attempting to solve the multi-access box

10) Age (adult: after hatch year, juvenile: hatch year). NOTE: this variable will be removed if only adults are tested (and we are planning to test only adults).

11) ID (random effect because multiple measures per individual)

12) Condition: control, flexibility manipulation

ANALYSIS 3 - *H1: P1 alternative 4: correlation between boldness and exploration*

Dependent variable: Boldness: Latency to land on the table - OR - Latency to eat the food - OR - Latency to touch a threatening object next to food (we will choose the variable with the most data)

Independent variables:

1) Time spent exploring the outside of the novel environment (within 20cm) before entering it

2) Latency to land on the table next to an object (novel, familiar) (that does not contain food) in a familiar environment (that contains maintenance diet away from the object) - OR - latency to touch an object (novel, familiar) (choose the variable with the most data)

ANALYSIS 4 - *H1: P3: does persistence correlate with reversal persistence?*

Dependent variable: The number of incorrect choices in the final reversal before making the first correct choice

Independent variables:

1) Average number of touches to the functional part of an apparatus per time (multi-access box, novel environment in P1, novel objects in P2 and P4)

2) Condition: control, flexibility manipulation

ANALYSIS 5 - *H2: P6: captive vs wild*

Dependent variables

1) Boldness: In captivity we will measure boldness as the latency to land on the table - OR - Latency to eat the food - OR - Latency to touch a threatening object that is next to food (we will choose the variable with the most data); In the wild the dependent variable will be the latency to come within 2m - OR - Latency to eat the food - OR - Latency to touch a threatening object that is next to food (we will choose the variable with the most data).

2) Persistence: Number of touches to an apparatus per time (multi-access box in the behavioral flexibility preregistration, novel environment in P1, objects in P2 and P4)

3) Exploration of novel environment: Latency to enter a novel sub-environment inside a familiar environment

4) Exploration of novel object: Latency to land next to an object (novel, familiar) (that does not contain food) in a familiar environment (that contains maintenance diet away from the object) - OR - latency to touch an object (novel, familiar) (choose the variable with the most data)

Note: if 3 and 4 are consistent within individuals, and correlate, we will combine these variables into one exploration propensity score.

Independent variables

1) Context: captive or wild

2) Number of times we attempted to assay boldness or exploration but failed due to lack of participation

3) ID (random effect because multiple measures per individual)

E. ANALYSIS PLAN

We do not plan to **exclude** any data. When **missing data** occur, the existing data for that individual will be included in the analyses for the tests they completed. Analyses will be conducted in R (current version 4.4.0; (R Core Team, 2023)). When there is more than one experimenter within a test, experimenter will be added as a random effect to account for potential differences between experimenters in conducting the tests. If there are no differences between models including or excluding experimenter as a random effect, then we will use the model without this random effect for simplicity.

Ability to detect actual effects To begin to understand what kinds of effect sizes we will be able to detect given our sample size limitations and our interest in decreasing noise by attempting to measure it, which increases the number of explanatory variables, we used G*Power (v.3.1, Faul et al. (2007), Faul et al. (2009)) to conduct power analyses based on confidence intervals. G*Power uses pre-set drop down menus and we chose the options that were as close to our analysis methods as possible (listed in each analysis below). Note that there were no explicit options for GLMs (though the chosen test in G*Power appears to align with GLMs) or GLMMs or for the inclusion of the number of trials per bird (which are generally large in our investigation), thus the power analyses are only an approximation of the kinds of effect sizes we can detect. We realize that these power analyses are not fully aligned with our study design and that these kinds of analyses are not appropriate for Bayesian statistics (e.g., our MCMCglmm below), however we are unaware of better options at this time. Additionally, it is difficult to run power analyses because it is unclear what kinds of effect sizes we should expect due to the lack of data on this species for these experiments.

To address the power analysis issues, we will run simulations on our Arizona data set before conducting any analyses in this preregistration. We will first run null models (i.e., dependent variable ~ 1 + random effects), which will allow us to determine what a weak versus a strong effect is for each model. Then we will run simulations based on the null model to explore the boundaries of influences (e.g., sample size) on our ability to detect effects of interest of varying strengths. If simulation results indicate that our Arizona sample size is not larger than the lower boundary, we will continue these experiments at the next field site until we meet the minimum suggested sample size.

Data checking The data will be checked for overdispersion, underdispersion, zero-inflation, and heteroscedasticity with the DHARMa R package (Hartig, 2019) following methods by Hartig. Note: DHARMa doesn't support MCMCglmm, therefore we will use the closest supported model: glmer from the R package lme4 (Bates et al., 2015).

Repeatability of exploration, boldness and persistence **Analysis:** We will obtain repeatability estimates that account for the observed and latent scales, and then compare them with the raw repeatability estimate from the null model. The repeatability estimate indicates how much of the total variance, after accounting for fixed and random effects, is explained by individual differences (ID). We will run this GLMM using the MCMCglmm function in the MCMCglmm package (Hadfield, 2010) with a Poisson distribution and log link using 13,000 iterations with a thinning interval of 10, a burnin of 3,000, and minimal priors (V=1, nu=0) (Hadfield, 2014). We will ensure the GLMM shows acceptable convergence (i.e., lag time autocorrelation values <0.01; (Hadfield, 2010)), and adjust parameters if necessary.

Note Feb 2021: a Gaussian distribution was used instead of a Poisson for exploration and boldness latencies because they are continuous variables.

Note: The power analysis is the same as for P3 (below) because there are the same number of explanatory variables (fixed effects).

Perhaps boldness is not repeatable because grackles are more likely to change their behavioral response to a potentially threatening object after the first exposure to that object. Consequently, this *unregistered post-hoc analysis* tests whether grackle boldness is repeatable across potentially threatening objects if we only consider their performance on the first trial.

H1: P1-P5: correlation of flexibility with exploration of new environments and objects, boldness, persistence, and motor diversity **Analysis:** If behavior is not repeatable across assays at Time 1 and Time 2 (six weeks apart, both assays occur after the flexibility manipulation takes place) for exploration, boldness, persistence, or motor diversity (see analysis for P6), we will not include these variables in analyses involving flexibility. If behavior is repeatable within individuals, we will examine the relationship between flexibility and these variables as follows. Note that the two exploration measures (novel environment and novel object) will be combined into one variable if they correlate and are both repeatable within individuals.

Because the independent variables could influence each other, we will analyze them in a single model: Generalized Linear Mixed Model (GLMM; MCMCglmm function, MCMCglmm package; (Hadfield, 2010)) with a Poisson distribution and log link using 13,000 iterations with a thinning interval of 10, a burnin of 3,000, and minimal priors (V=1, nu=0) (Hadfield, 2014). We will ensure the GLMM shows acceptable convergence (i.e., lag time autocorrelation values <0.01; (Hadfield, 2010)), and adjust parameters if necessary. We will determine whether an independent variable had an effect or not using the Estimate in the full model.

To roughly estimate our ability to detect actual effects (because these power analyses are designed for frequentist statistics, not Bayesian statistics), we ran a power analysis in G*Power with the following settings: test family=F tests, statistical test=linear multiple regression: Fixed model (R^2 deviation from zero), type of power analysis=a priori, alpha error probability=0.05. We reduced the power to 0.70 and increased the effect size until the total sample size in the output matched our projected sample size (n=32). The number of predictor variables was restricted to only the fixed effects because this test was not designed for mixed models. The protocol of the power analysis is here:

Input:

Effect size f^2 = 0,62

err prob = 0,05

Power (1- err prob - note: =probability of making a Type II error) = 0,7

Number of predictors = 10

Output:

Noncentrality parameter = 19,8400000

Critical F = 2,3209534

Numerator df = 10

Denominator df = 21

Total sample size = 32

Actual power = 0,7027626

This means that, with our sample size of 32, we have a 70% chance of detecting a large effect (approximated at f^2=0.35 by Cohen (1988)).

H1: P1-P5 alternative: Control vs flexibility manipulated individuals The flexibility manipulation did work such that individuals in the serial reversal learning group increased their speed to pass each reversal. After we received in-principal recommendation for the preregistration associated with this research, we developed and tested the flexibility comprehensive variable. We found that this variable more accurately represented flexible behavior (Blaisdell et al., 2021; Lukas et al., 2022). However, our preregistered predictions still included comparison of performance on the behavioral trait assays between control and manipulated individuals. Thus, we conducted these comparisons, above in the post-study manuscript. *NOTE that we preregistered that we would run this analysis, but we did not preregister any code*

H1: P1 alternative 4: correlations between exploration and boldness measures **Analysis:** Generalized Linear Model (GLM; glm function, stats package) with a Poisson distribution and log link. For an estimation of our ability to detect actual effects, please see the power analysis for P3 below.

Model validation: Determine whether the test model results are likely to be reliable given the data (Burnham & Anderson, 2003). Compare Akaike weights (range: 0–1, the sum of all model weights equals 1; Akaike, 1981) between the test model and a base model (number of trials to reverse as the response variable and 1 as the explanatory variable) using the dredge function in the MuMIn package (Bates et al., 2012). If the best fitting model has a high Akaike weight (>0.89; (Burnham & Anderson, 2003)), then it indicates that the results are likely given the data. The Akaike weights indicate the best fitting model is the [base/test - *delete as appropriate*] model (Table 2).

H1: P3: correlations between persistence measures **Analysis:** Generalized Linear Model (GLM; glm function, stats package) with a Poisson distribution and log link.

To determine our ability to detect actual effects, we ran a power analysis in G*Power with the following settings: test family=F tests, statistical test=linear multiple regression: Fixed model (R^2 deviation from zero), type of power analysis=a priori, alpha error probability=0.05. We reduced the power to 0.70 and increased the effect size until the total sample size in the output matched our projected sample size (n=32). The protocol of the power analysis is here:

Input:

Effect size f^2 = 0,27

err prob = 0,05

Power (1- err prob - note: =probability of making a Type II error) = 0,7

Number of predictors = 2

Output:

Noncentrality parameter = 8,6400000

Critical F = 3,3276545

Numerator df = 2

Denominator df = 29

Total sample size = 32

Actual power = 0,7047420

This means that, with our sample size of 32, we have a 70% chance of detecting a medium (approximated at f^2=0.15 by Cohen (1988)) to large effect (approximated at f^2=0.35 by Cohen (1988)).

Model validation: Determine whether the test model results are likely to be reliable given the data (Burnham & Anderson, 2003). Compare Akaike weights (range: 0–1, the sum of all model weights equals 1; Akaike, 1981) between the test model and a base model (number of trials to reverse as the response variable and 1 as the explanatory variable) using the dredge function in the MuMIn package (Bates et al., 2012). If the best fitting model has a high Akaike weight (>0.89; (Burnham & Anderson, 2003)), then it indicates that the results are likely given the data. The Akaike weights indicate the best fitting model is the [base/test - *delete as appropriate*] model (Table 2).

H2: P6: captive vs wild A GLMM (as in the repeatability analysis) will be conducted.

Alternative Analyses We anticipate that we will want to run additional/different analyses after reading McElreath (2016). We will revise this preregistration to include these new analyses before conducting the analyses above.

F. ETHICS

This research is carried out in accordance with permits from the:

1) US Fish and Wildlife Service (scientific collecting permit number MB76700A-0,1,2)
2) US Geological Survey Bird Banding Laboratory (federal bird banding permit number 23872)
3) Arizona Game and Fish Department (scientific collecting license number SP594338 [2017] and SP606267 [2018])
4) Institutional Animal Care and Use Committee at Arizona State University (protocol number 17-1594R)
5) University of Cambridge ethical review process (non-regulated use of animals in scientific procedures: zoo4/17)

G. AUTHOR CONTRIBUTIONS

McCune: Hypothesis development, data collection, data analysis and interpretation, write up, revising/editing.

MacPherson: Data collection, data interpretation, revising/editing.

Rowney: Data collection, data interpretation, revising/editing.

Bergeron: Data collection, data interpretation, revising/editing.

Folsom: Data collection, data interpretation, revising/editing.

Deffner: Data analysis (Flexibility comprehensive model), data interpretation, revising/editing.

Logan: Hypothesis development, data collection, data analysis and interpretation, revising/editing, materials/funding.

H. FUNDING

This research is funded by the Department of Human Behavior, Ecology and Culture at the Max Planck Institute for Evolutionary Anthropology, and by a Leverhulme Early Career Research Fellowship to Logan in 2017-2018.

###I. CONFLICT OF INTEREST DISCLOSURE

We, the authors, declare that we have no financial conflicts of interest with the content of this article. Corina Logan is a Recommender and on the Managing Board at PCI Ecology.

J. ACKNOWLEDGEMENTS

We thank Dieter Lukas for help polishing the predictions; Ben Trumble for providing us with a wet lab at Arizona State University and Angela Bond for lab support; Melissa Wilson for sponsoring our affiliations at Arizona State University and lending lab equipment; Kevin Langergraber for serving as local PI on the ASU IACUC; Kristine Johnson for technical advice on great-tailed grackles; Arizona State University School of Life Sciences Department Animal Care and Technologies for providing space for our aviaries and for their excellent support of our daily activities; Julia Cissewski for tirelessly solving problems involving financial transactions and contracts; Richard McElreath for project support; Aaron Blackwell and Ken Kosik for being the UCSB sponsors of the Cooperation Agreement with the Max Planck Institute for Evolutionary Anthropology; Jeremy Van Cleve, our Recommender at PCI Ecology, and two anonymous reviewers for their wonderful feedback; Vincent Kiepsch and Sierra Planck for interobserver reliability video coding: Sawyer Lung for field support; Alexis Breen for video coding; and our research assistants: Aelin Mayer, Nancy Rodriguez, Brianna Thomas, Aldora Messinger, Elysia Mamola, Michael Guillen, Rita Barakat, Adriana Boderash, Olateju Ojekunle, August Sevchik, Justin Huynh, Jennifer Berens, Amanda Overholt, Michael

Pickett, Sam Munoz, Sam Bowser, Emily Blackwell, Kaylee Delcid, Sofija Savic, Brynna Hood, Sierra Planck, and Elise Lange.

K. REFERENCES

Amy M, Van Oers K, Naguib M (2012) Worms under cover: Relationships between performance in learning tasks and personality in great tits (parus major). *Animal Cognition*, **15**, 763–770.

Audet J-N, Couture M, Lefebvre L, Jarvis ED (2024) Problem-solving skills are predicted by technical innovations in the wild and brain size in passerines. *Nature Ecology & Evolution*, **8**, 806–816.

Bates D, Mächler M, Bolker B, Walker S (2015) Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, **67**, 1–48. https://doi.org/10.18637/jss.v067.i01

Bates D, Maechler M, Bolker B (2012) lme4: Linear mixed-effects models using S4 classes (2011). R package version 0.999375-42.

Bebus SE, Small TW, Jones BC, Elderbrock EK, Schoech SJ (2016) Associative learning is inversely related to reversal learning and varies with nestling corticosterone exposure. *Animal Behaviour*, **111**, 251–260.

Bensky MK, Bell AM (2022) A behavioral syndrome linking boldness and flexibility facilitates invasion success in sticklebacks. *The American Naturalist*, **200**, 846–856.

Blaisdell A, Seitz B, Rowney C, Folsom M, MacPherson M, Deffner D, Logan CJ (2021) Do the more flexible individuals rely more on causal cognition? Observation versus intervention in causal inference in great-tailed grackles. *Peer Community Journal*, **1**. https://doi.org/10.24072/pcjournal.44

Burnham KP, Anderson DR (2003) *Model selection and multimodel inference: A practical information-theoretic approach.* Springer Science & Business Media.

Canestrelli D, Bisconti R, Carere C (2016) Bolder takes all? The behavioral dimension of biogeography. *Trends in ecology & evolution*, **31**, 35–43.

Carter AJ, Feeney WE, Marshall HH, Cowlishaw G, Heinsohn R (2013) Animal personality: What are behavioural ecologists measuring? *Biological Reviews*, **88**, 465–475.

Chapple DG, Simmonds SM, Wong BB (2012) Can behavioral and personality traits influence the success of unintentional species introductions? *Trends in ecology & evolution*, **27**, 57–64. https://doi.org/10.1016/j.tree.2011.09.010

Chow PKY, Lea SE, Leaver LA (2016) How practice makes perfect: The role of persistence, flexibility and learning in problem-solving efficiency. *Animal behaviour*, **112**, 273–283. https://doi.org/10.1016/j.anbehav.2015.11.014

Cohen J (1988) Statistical power analysis for the behavioral sciences 2nd edn.

Cohen JD, McClure SM, Yu AJ (2007) Should i stay or should i go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **362**, 933–942.

Coleman SL, Mellgren RL (1994) Neophobia when feeding alone or in flocks in zebra finches, taeniopygia guttata. *Animal Behaviour*, **48**, 903–907.

De Meester G, Pafilis P, Van Damme R (2022) Bold and bright: Shy and supple? The effect of habitat type on personality–cognition covariance in the aegean wall lizard (podarcis erhardii). *Animal Cognition*, 1–23.

Dingemanse NJ, Dochtermann NA (2013) Quantifying individual variation in behaviour: Mixed-effect modelling approaches. *Journal of Animal Ecology*, **82**, 39–54.

Diquelou MC, Griffin AS, Sol D (2015) The role of motor diversity in foraging innovations: A cross-species comparison in urban birds. *Behavioral Ecology*, **27**, 584–591. https://doi.org/10.1093/beheco/arv190

Dougherty LR, Guillette LM (2018) Linking personality and cognition: A meta-analysis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **373**, 20170282.

Duckworth RA (2010) Evolution of personality: Developmental constraints on behavioral flexibility. *The Auk*, **127**, 752–758.

Faul F, Erdfelder E, Buchner A, Lang A-G (2009) Statistical power analyses using g* power 3.1: Tests for correlation and regression analyses. *Behavior research methods*, **41**, 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Faul F, Erdfelder E, Lang A-G, Buchner A (2007) G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*, **39**, 175–191. https://doi.org/10.3758/BF03193146

Fidler AE, Oers K van, Drent PJ, Kuhn S, Mueller JC, Kempenaers B (2007) Drd4 gene polymorphisms are associated with personality variation in a passerine bird. *Proceedings of the Royal Society B: Biological Sciences*, **274**, 1685–1691.

Gamer M, Lemon J, Gamer MM, Robinson A, Kendall's W (2012) Package "irr." *Various coefficients of interrater reliability and agreement.*

Greggor AL, Thornton A, Clayton NS (2015) Neophobia is not only avoidance: Improving neophobia tests by combining cognition and ecology. *Current Opinion in Behavioral Sciences*, **6**, 82–89.

Griffin AS, Diquelou MC (2015) Innovative problem solving in birds: A cross-species comparison of two highly successful passerines. *Animal Behaviour*, **100**, 84–94. https://doi.org/10.1016/j.anbehav.2014.11.012

Griffin AS, Guez D, Federspiel I, Diquelou M, Lermite F (2016) Invading new environments: A mechanistic framework linking motor diversity and cognition to establishment success. *Biological Invasions and Animal Behaviour*, 26e46. https://doi.org/10.1017/CBO9781139939492.004

Guenther A, Brust V, Dersen M, Trillmich F (2014) Learning and personality types are related in cavies (cavia aperea). *Journal of Comparative Psychology*, **128**, 74.

Hadfield J (2010) MCMC methods for multi-response generalized linear mixed models: The MCMCglmm r package. *Journal of Statistical Software*, **33**, 1–22. https://doi.org/10.18637/jss.v033.i02

Hadfield J (2014) MCMCglmm course notes.

Hartig F (2019) *DHARMa: Residual diagnostics for hierarchical (multi-level / mixed) regression models.*

Hendry AP, Farrugia TJ, Kinnison MT (2008) Human influences on rates of phenotypic change in wild animal populations. *Molecular ecology*, **17**, 20–29.

Herborn KA, Heidinger BJ, Alexander L, Arnold KE (2014) Personality predicts behavioral flexibility in a fluctuating, natural environment. *Behavioral Ecology*, **25**, 1374–1379.

Homberg JR, Pattij T, Janssen MC, Ronken E, De Boer SF, Schoffelmeer AN, Cuppen E (2007) Serotonin transporter deficiency in rats improves inhibitory control but not behavioural flexibility. *European Journal of Neuroscience*, **26**, 2066–2073. https://doi.org/10.1111/j.1460-9568.2007.05839.x

Hutcheon JA, Chiolero A, Hanley JA (2010) Random measurement error and regression dilution bias. *Bmj*, **340**, c2289. https://doi.org/10.1136/bmj.c2289

Johnson K, Peer BD (2001) *Great-tailed grackle: Quiscalus mexicanus.* Birds of North America, Incorporated.

Logan C (2016a) How far will a behaviourally flexible invasive bird go to innovate? *Royal Society Open Science*, **3**, 160247.

Logan CJ (2016b) Behavioral flexibility in an invasive bird is independent of other behaviors. *PeerJ*, **4**, e2215. https://doi.org/10.7717/peerj.2215

Logan CJ (2016c) Behavioral flexibility and problem solving in an invasive bird. *PeerJ*, **4**, e1975. https://doi.org/10.7717/peerj.1975

Logan C, Lukas D, Blaisdell A, Johnson-Ulrich Z, MacPherson M, Seitz B, Sevchik A, McCune K (2023) Behavioral flexibility is manipulable and it improves flexibility and innovativeness in a new context. *Peer Community Journal*, **3**. https://doi.org/10.24072/pcjournal.284

Logan C, Lukas D, Geng X, LeGrande-Rolls C, Marfori Z, MacPherson M, Rowney C, Smith C, McCune K (2024) Behavioral flexibility is related to foraging, but not social or habitat use behaviors in a species that is rapidly expanding its range. *EcoEvoRxiv.* https://doi.org/10.32942/X2T036

Lukas D, McCune K, Blaisdell A, Johnson-Ulrich Z, MacPherson M, Seitz B, Sevchik A, Logan C (2022) Behavioral flexibility is manipulatable and it improves flexibility and problem solving in a new context: Post-hoc analyses of the components of behavioral flexibility. *EcoEvoRxiv.* https://doi.org/10.32942/osf.io/4ycps

McCune K, Blaisdell A, Johnson-Ulrich Z, Sevchik A, Lukas D, MacPherson M, Seitz B, Logan C (2023) Repeatability of performance within and across contexts measuring behavioral flexibility. *PeerJ.* https://doi.org/10.7717/peerj.15773

McCune K, Jablonski P, Lee S, Ha R (2018) Evidence for personality conformity, not social niche specialization in social jays. *Behavioral Ecology*, **29**, 910–917.

McElreath R (2016) *Statistical rethinking: A bayesian course with examples in r and stan.* CRC Press. https://doi.org/10.1201/9781315372495

Mettke-Hofmann C, Lorentzen S, Schlicht E, Schneider J, Werner F (2009) Spatial neophilia and spatial neophobia in resident and migratory warblers (sylvia). *Ethology*, **115**, 482–492. https://doi.org/10.1111/j.1439-0310.2009.01632.x

Mettke-Hofmann C, Winkler H, Leisler B (2002) The significance of ecological factors for exploration and neophobia in parrots. *Ethology*, **108**, 249–272. https://doi.org/10.1046/j.1439-0310.2002.00773.x

Mikhalevich I, Powell R, Logan C (2017) Is behavioural flexibility evidence of cognitive complexity? How evolution can inform comparative cognition. *Interface Focus*, **7**, 20160121. https://doi.org/10.1098/rsfs.2016.0121

Morand-Ferron J, Overington S, Cauchard L, Lefebvre L (2009) Innovation in groups: Does the proximity of others facilitate or inhibit performance? *Behaviour*, **146**, 1543–1564.

Morand-Ferron J, Reichert MS, Quinn JL (2022) Cognitive flexibility in the wild: Individual differences in reversal learning are explained primarily by proactive interference, not by sampling strategies, in two passerine bird species. *Learning & Behavior*, **50**, 153–166. https://doi.org/10.3758/s13420-021-00505-1

Perals D, Griffin AS, Bartomeus I, Sol D (2017) Revisiting the open-field test: What does it really tell us about animal personality? *Animal Behaviour*, **123**, 69–79.

Plummer M, Best N, Cowles K, Vines D K, Bates D, Almond R, Magnusson A (2020) Coda: Output analysis and diagnostics for MCMC (2020). R package version 2.14.0.

R Core Team (2023) *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Réale D, Reader SM, Sol D, McDougall PT, Dingemanse NJ (2007) Integrating animal temperament within ecology and evolution. *Biological reviews*, **82**, 291–318. https://doi.org/10.1111/j.1469-185X.2007.00010.x

Rojas-Ferrer I, Thompson MJ, Morand-Ferron J (2020) Is exploration a metric for information gathering? Attraction to novelty and plasticity in black-capped chickadees. *Ethology*, **126**, 383–392.

Rowe C, Healy SD (2014) Measuring variation in cognition. *Behavioral Ecology*, **25**, 1287–1292.

Schuett W, Dall SR (2009) Sex differences, social context and personality in zebra finches, taeniopygia guttata. *Animal Behaviour*, **77**, 1041–1050.

Shaw RC, Schmelz M (2017) Cognitive test batteries in animal cognition research: Evaluating the past, present and future of comparative psychometrics. *Animal cognition*, **20**, 1003–1018.

Sih A (2013) Understanding variation in behavioural responses to human-induced rapid environmental change: A conceptual overview. *Animal Behaviour*, **85**, 1077–1088.

Sih A, Bell A, Johnson JC (2004) Behavioral syndromes: An ecological and evolutionary overview. *Trends in ecology & evolution*, **19**, 372–378.

Sih A, Ferrari MC, Harris DJ (2011) Evolution and behavioural responses to human-induced rapid environmental change. *Evolutionary applications*, **4**, 367–387.

Sol D, Lapiedra O, González-Lagos C (2013) Behavioural adjustments for a life in the city. *Animal behaviour*, **85**, 1101–1112. https://doi.org/10.1016/j.anbehav.2013.01.023

Sol D, Timmermans S, Lefebvre L (2002) Behavioural flexibility and invasion success in birds. *Animal behaviour*, **63**, 495–502. https://doi.org/10.1006/anbe.2001.1953

Stoffel MA, Nakagawa S, Schielzeth H (2017) rptR: Repeatability estimation and variance decomposition by generalized linear mixed-effects models. *Methods in Ecology and Evolution*, **8**, 1639–1644. https://doi.org/10.1111/2041-210X.12797

Summers J, Lukas D, Logan C, Chen N (2023) The role of climate change and niche shifts in divergent range dynamics of a sister-species pair. *Peer Community Journal*, **3**. https://doi.org/10.24072/pcjournal.248

Takola E, Krause ET, Müller C, Schielzeth H (2021) Novelty at second glance: A critical appraisal of the novel object paradigm based on meta-analysis. *Animal Behaviour*, **180**, 123–142.

Titulaer M, Oers K van, Naguib M (2012) Personality affects learning performance in difficult tasks in a sex-dependent way. *Animal Behaviour*, **83**, 723–730.

Wehtje W (2003) The range expansion of the great-tailed grackle (quiscalus mexicanus gmelin) in north america since 1880. *Journal of Biogeography*, **30**, 1593–1607. https://doi.org/10.1046/j.1365-2699.2003.00970.x

Wright TF, Eberhard JR, Hobson EA, Avery ML, Russello MA (2010) Behavioral flexibility and species invasions: The adaptive flexibility hypothesis. *Ethology Ecology & Evolution*, **22**, 393–404. https://doi.org/10.1080/03949370.2010.505580

Zhang X, Yi N (2020) NBZIMM: Negative binomial and zero-inflated mixed models, with application to microbiome/metagenomics data analysis. *BMC bioinformatics*, **21**, 1–19.