Cover letter

Dear PCI Ecology,

We are submitting one of the articles that is resulting from our in principle recommended preregistration titled “Implementing a rapid geographic range expansion - the role of behavior and habitat changes”. The current article addresses Question 1 (behavior in great-tailed grackles), therefore we removed the parts of the preregistration that did not pertain to this question. We also moved most of the methods sections to supplementary material for ease of reading. Changes are shown in the tracked changes version at the PCI Ecology website and at the Rmd file at GitHub: https://github.com/corinalogan/grackles/blob/master/Files/Preregistrations/gxpopbehaviorhabitatq1.Rmd . Note that the addition of the Results and Discussion sections are not highlighted in track changes in the PCI Ecology version because these sections are entirely new.

There are two changes in particular that we would like to bring to your attention:

1) We added a description of the exploration assay to “Methods > Protocols and open materials” because we realized we forgot to include this in the preregistration.

2) We originally stated that we would run interobserver reliability analyses on at least 20% of the videos. We are above 20% for the reversal and innovation measures, however the exploration measure ended up having a larger sample size than we expected. As such, we only have interobserver reliability data for 15% of the exploration videos. When researchers conduct interobserver reliability, they generally code 10-20% of the videos, so our 15% for exploration is well within this range. We would like to stay with what we have because exploration in this and a previous study was highly repeatable. If we were to bring the number of videos coded to 20%, it would require several months to complete the work because our only trained video coder will not be available until summer 2023. If you deem it important to bring this number to 20%, please let us know and we will have the coder conduct this work and incorporate the updated interobserver reliability number in a revised version.

In summary, this is the POST-study (Stage 2) manuscript of a preregistration that received in principle acceptance, which means that the introduction, hypothesis, and methods have previously passed peer review at PCI Ecology. As such, this review should focus on the results and discussion sections, which are new. For information on how to recommend and review Stage 2 manuscripts, please refer to the PCI Registered Reports guidelines for recommenders (https://rr.peercommunityin.org/help/guide_for_recommenders#h_675964623640161364339090 5) and reviewers (https://rr.peercommunityin.org/help/guide_for_reviewers). Ideally, the same recommender and reviewers for the preregistration would come back for this post-study manuscript. We are not in a hurry so please feel free to take your time with the manuscript to accommodate recommender and reviewer schedules.

We thank you very much for considering this article!

All our best,
Corina (on behalf of all co-authors)

Implementing a rapid geographic range expansion - the role of behavior ~~and habitat~~ changes

Logan CJ[1]* McCune KB[2] LeGrande-Rolls C[1] Marfori Z[1] Hubbard J[3] ~~Chen N[3]~~ Lukas D[1]

Affiliations: 1) Max Planck Institute for Evolutionary Anthropology, 2) University of California Santa Barbara, 3) Animal Behavior Graduate Group, University of California Davis~~Rochester~~.
*Corresponding author: corina_logan@eva.mpg.de

ABSTRACT

It is generally thought that behavioral flexibility, the ability to change behavior when circumstances change, plays an important role in the ability of ~~a~~ species to rapidly expand their geographic range~~ (Chow et al., 2016; Griffin & Guez, 2014; e.g. Lefebvre et al., 1997; Sol et al., 2002, 2005, 2007; Sol & Lefebvre, 2000). However, it is an alternative non-exclusive possibility that an increase in the amount of available habitat can also facilitate a range expansion (Hanski & Gilpin, 1991; Wiens, 1997).~~ Great-tailed grackles (*Quiscalus mexicanus*) are a social, polygamous species that is rapidly expanding its geographic range ~~(Wehtje, 2003)~~ by settling in new areas and habitats ~~(@summers2023xpop). and eats a variety of human foods in addition to foraging on insects and on the ground for other natural food items (K. Johnson & Peer, 2001).~~ They are behaviorally flexible ~~(Logan, 2016a)~~ and highly associated with human-modified environments ~~(K. Johnson & Peer, 2001)~~, eating a variety of human foods in addition to foraging on insects and on the ground for other natural food items ~~(K. Johnson & Peer, 2001)~~. They~~,~~ ~~thus~~ offer~~ing~~ an opportunity to assess the role of behavior ~~and habitat~~ change over the course of their expansion. We ~~first aim to~~ compare behavior in wild-caught grackles from two~~three~~ populations across their range (~~core of the original range,~~ an older~~more recent~~ population in the middle of the northern expansion front: Tempe, Arizona, and a ~~very~~ more recent population on the northern edge of the expansion front: Woodland, California) to investigate whether~~:~~ ~~1)~~ certain behaviors (flexibility, innovativeness, exploration, and persistence) have higher averages and variances in the newer~~edge~~ or older ~~the middle~~some populations ~~relative to others~~, ~~, and 2) individuals in a more recently established population exhibit more dispersal behavior (i.e., individuals are more likely to move away from their parents). Secondly, we aim to investigate whether habitat availability, not necessarily inherent species differences, can explain why great-tailed grackles are able to much more rapidly expand their range than their closest relative, boat-tailed grackles (*Q. major*) (Post et al., 1996; Wehtje, 2003). We will examine temporal habitat changes over the past few decades using existing databases on presence/absence of both grackle species and compare habitat variables to determine whether: 3) these species use different habitats, habitat suitability and connectivity (which combined determines whether habitat is available) has increased across their range, and what proportion of suitable habitat both species occupy. **Finally, we will 4) determine whether changes in behavioral traits facilitate the rapid GTGR expansion by comparing their behavior with BTGR on the same tests in aim 1.** We find that grackles in the edge population were more innovative and less exploratory, and that there were no population differences in flexibility (measured by reversal learning) or persistence (the proportion of trials

participated in). Results will elucidate that whether the rapid geographic range expansion of great-tailed grackles is associated with individuals differentially expressing particular behaviors and/or whether the expansion is facilitated by the alignment of their natural behaviors with an increase in suitable habitat (i.e., human-modified environments). Our findings highlight the value of population studies and of breaking down cognitive concepts into direct measures of individual abilities to better understand how species might adapt to novel circumstances.

INTRODUCTION

It is generally thought that behavioral flexibility, the ability to change behavior when circumstances change through packaging information and making it available to other cognitive processes (see Mikhalevich et al., 2017 for theoretical background on our flexibility definition), plays an important role in the ability of a species to rapidly expand their geographic range (Chow et al., 2016; Griffin & Guez, 2014; e.g., Lefebvre et al., 1997; Sol et al., 2002, 2005, 2007; Sol & Lefebvre, 2000). These ideas predict that flexibility, exploration, and innovation [creating new behaviors or using existing behaviors in a new context, @griffin2014innovation] facilitate the expansion of individuals into completely new areas and that their role diminishes after a certain number of generations (Wright et al., 2010). In support of this, experimental studies have shown that latent abilities are primarily expressed in a time of need (A. M. Auersperg et al., 2012; Bird & Emery, 2009; Laumer et al., 2018; Manrique & Call, 2011; e.g., Taylor et al., 2007). Therefore, we do not expect the founding individuals who initially dispersed out of their original range to have unique behavioral characteristics that are passed on to their offspring. Instead, we expect that the actual act of continuing a range expansion relies on flexibility, exploration, innovation, and persistence, and that these behaviors are therefore expressed more on the edge of the expansion range where there have not been many generations to accumulate relevant knowledge about the environment.

It is also possible that a recent increase in the amount of available habitat can facilitate a geographic range expansion (Hanski & Gilpin, 1991; Wiens, 1997). A species may not need to be behaviorally flexible to move into new areas if they can continue to use the same types of habitat they are accustomed to. Human-modified environments are increasing (Goldewijk, 2001; e.g., Liu et al., 2020; Wu et al., 2011), and species associated with these habitats show differences in their behavior (Chejanovski et al., 2017; e.g., Ciani, 1986; Federspiel et al., 2017). These species offer an opportunity for simultaneous investigation of the roles of behavior and increased habitat availability for a rapidly increasing geographic range expansion.

To determine whether behavior is involved in a rapid geographic range expansion, direct measures of individual behavioral abilities must be collected in populations across the range of the species (see the discussion on the danger of proxies of flexibility in Logan et al., 2018). Our study aimsWe planned to test whether behavioral flexibility and/or an increase in habitat availability playsed a role in the rapid geographic range expansion of great-tailed grackles (*Quiscalus mexicanus*). Great-tailed grackles are behaviorally flexible (Logan, 2016a), rapidly expanding their geographic range (Wehtje, 2003), and highly associated with human-modified environments (K. Johnson & Peer, 2001), thus offering an opportunity to assess the role of behavior and habitat changes over the course of their expansion. This social, polygamous species eats a variety of human foods in addition to foraging on insects and on the ground for other natural food items (K. Johnson & Peer, 2001). This feature increases the ecological relevance of comparative cognition experiments that measure individual behavior abilities: grackles eat at outdoor cafes, from garbage cans, and onthey eat our crops. As such, they generally gain experience in the wild with approaching and opening novel objects to seek food (e.g., attempting to open a ketchup packet at an outdoor cafe, climbing into garbage cans to get french fries at the zoo, dunking sugar packets in water), which makes the tests involving human-made apparatuses ecologically relevant for this species.

We first aim to compare behavior in wild-caught great-tailed grackles from twothree populations across their range (core of the original range: Central America, an oldermore recent population in the middle of the northern expansion front: Tempe, Arizona using previously(for which we use previously published data from @logan2023flexmanip), and a moreevery recent population on the northern edge of the expansion front: Woodland, California) (Figure 1, Table 1). We will investigate whether certain behaviors have higher averages and variances in the edge population relative to the older populations. Specifically, we will investigate behavioral flexibility, measured as reversal learning of food-filled colored tube preferences (Logan, 2016b; Logan et al., 202319);, innovativeness, measured as the number of loci they solve to access food on a puzzle box (A. M. I. Auersperg et al.,

2011; Logan et al., 20232319);; exploration, measured as the latency to approach a novel environmentobject in the absence of nearby food (McCune KB et al., 2019; Mettke-Hofmann et al., 2009);; and persistence, measured as the proportion of sessions they participated in during the flexibility and innovativeness experiments (Figure 2). We will also examine whether individuals in a recently established population (California) are more likely to move away from the location they hatched by determining whether their average relatedness (calculated using single nucleotide polymorphisms, SNPs) is lower than what we would expect if individuals move randomly (Sevchik et al., 2019).

Second, we aim to investigate whether habitat availability, not necessarily inherent species differences, explains why great-tailed grackles are able to much more rapidly expand their range than their closest relative, boat-tailed grackles (*Q. major*) (Post et al., 1996; Wehtje, 2003). Detailed reports on the breeding ecology of these two species indicate that range expansion in boat- but not great-tailed grackles may be constrained by the availability of suitable nesting sites (Selander & Giller, 1961; Wehtje, 2003). Boat-tailed grackles nest primarily in coastal marshes, whereas great-tailed grackles nest in a variety of locations (e.g., palm trees, bamboo stalks, riparian vegetation, pines, oaks). However, this apparent difference in habitat breadth has yet to be rigorously quantified. Great-tailed grackles inhabit a wide variety of habitats (but not forests) at a variety of elevations (0-2134m), while remaining near water bodies, while boat-tailed grackles exist mainly in coastal areas (Selander & Giller, 1961). Both species have similar foraging habits: they are generalists and forage in a variety of substrates on a variety of different food items (Selander & Giller, 1961). We will use ecological niche modeling to examine temporal habitat changes over the past few decades using observation data for both grackle species from existing citizen science databases. We will compare this data with existing data on a variety of habitat variables. We identified suitable habitat variables from Selander & Giller (1961), K. Johnson & Peer (2001), and Post et al. (1996) (e.g., types of suitable land cover including marine coastal, wetlands, arable land, grassland, mangrove, urban), and we added additional variables relevant to our hypotheses (e.g., distance to nearest uninhabited suitable habitat patch to the north, presence/absence of water in the area). A suitable habitat map will be generated across the Americas using ecological niche models. This will allow us to determine whether the range of great-tailed grackles, but not boat-tailed grackles, might have increased because their habitat suitability and connectivity (which combined determines whether habitat is available) has increased, or whether great-tailed grackles now occupy a larger proportion of habitat that was previously available.

Third, we aim to compare behavior in one population of wild-caught boat-tailed grackles with that of great-tailed grackles on the same tests in aim 1 (behavioral flexibility, innovativeness, exploration, and persistence, but not dispersal). Similar to great-tailed grackles, boat-tailed grackles are social and polygamous, and eat human foods (Post, 1992; Post et al., 1996), which increases the ecological relevance of these tests. Determining whether great-tailed grackles perform better on these tests would provide support for the hypothesis that their behavior could be causing their rapid geographic range expansion. Alternatively, if boat-tailed and great-tailed grackles perform similarly, this would suggest that environmental, rather than behavioral, variables may play a larger role in restricting the boat-tailed grackle range expansion.

There could be multiple mechanisms underpinning the results we find, however our aim is to narrow down the role of changes in behavior and changes in habitats in the range expansion of great-tailed grackles. Our rResults will elucidate whether demonstrate that the rapid geographic range expansion of great-tailed grackles is associated with individuals differentially expressing particular behaviors in the edge compared to the olderestablished population and/or whether the expansion is facilitated by the alignment of their natural behaviors with an increase in suitable habitat (i.e., human-modified environments).

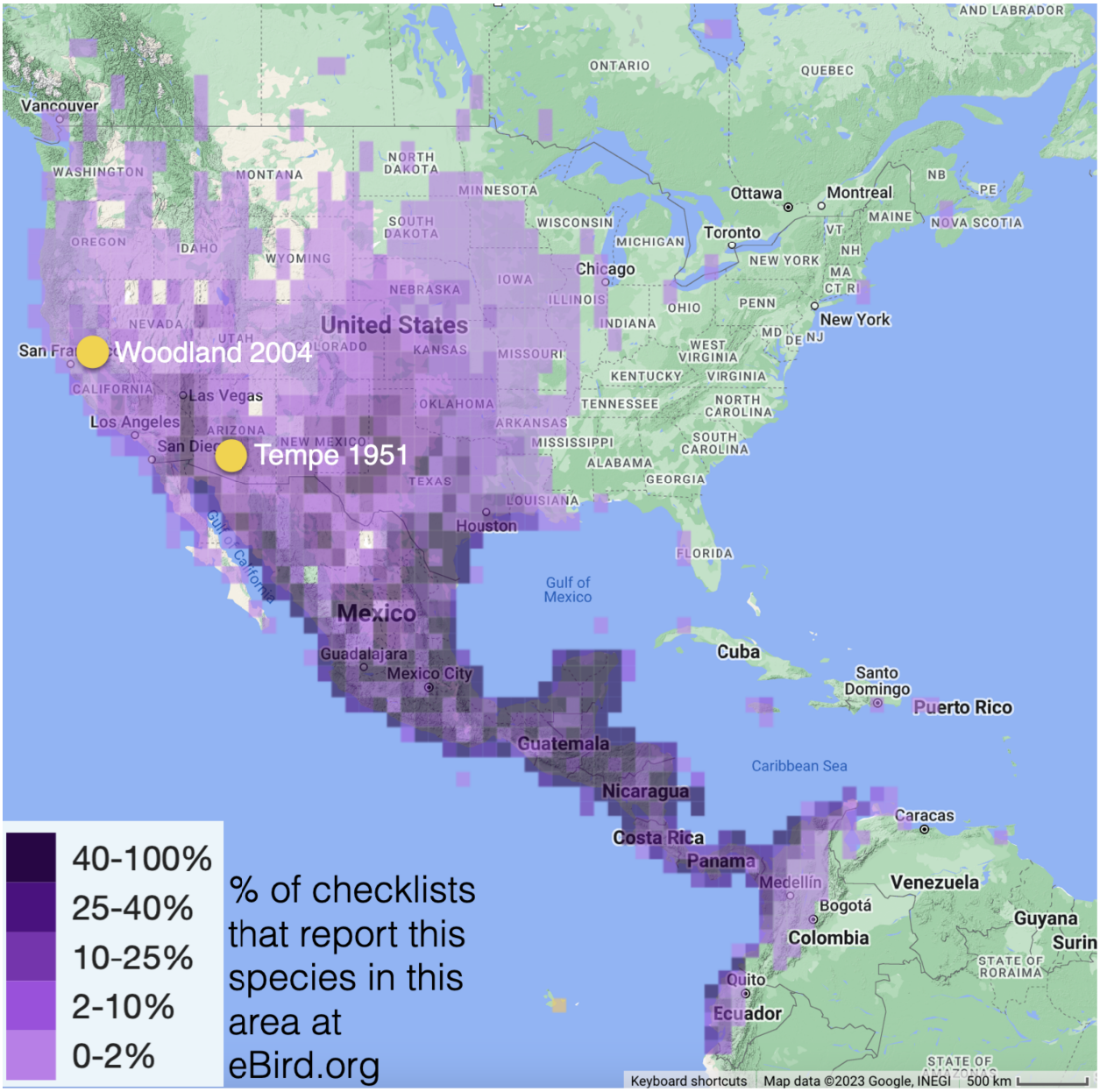


Figure 1. Great-tailed grackle field sites: Woodland is ~~a~~are recently established population~~s~~ (first breeding at the trapping location recorded in 2004) on the northern edge of the range, and Tempe is an older population (established in 1951) in the middle of the northern expansion front. Data from eBird.org

**Table 1. Population characteristics for the field sites. The number of generations at a site is based on a generation length of 5.6 years for this species [International (2018); note that this species starts breeding at age 1]~~,~~ and on the first year in which this species was reported (or estimated) to breed at each location (Woodland, California: Yolo Audubon Society's newsletter *The Burrowing Owl* from ~~(~~July 2004~~), which Steve Hampton shared with Logan~~; and Tempe, Arizona: estimated based on 1945 first-sighting report in nearby Phoenix, Arizona (Wehtje, 2004) to which we added 6 years~~,~~ to account for ~~which is~~ the average time between first-sighting and first-breeding - see Table 3 in ~~(~~Wehtje, 2003~~)~~. The average number of generations was calculated up to 2020, the final year of data collection in Tempe, and 2022, the final year of data collection in Woodland.

Site	Range position	Breeding since	Number of years breeding	Average number of generations	Citation
Tempe, Arizona	Middle of expansion	1951	69	12.3	Wehtje 2003, 2004
Woodland, California	Northern edge	2004	18	3.2	Burrowing Owl July 2004, Pandolfino et al. 2009

~~### A. STATE OF THE DATA~~

~~This preregistration was written (Mar 2020) prior to collecting any data from the edge and core populations, therefore we were blind to these data. However, we were not blind to some of the data from the Arizona population: some of the relatedness data (SNPs used for Hypothesis 2 to quantify relatedness to infer whether individuals disperse away from relatives) from the middle population (Arizona) has already been analyzed for other purposes (n=57 individuals, [see](http://corinalogan.com/Preregistrations/gdispersal_manuscript.html) @sevchik2019dispersal). Therefore, it will be considered secondary data: data that are in the process of being collected for other investigations. We have now collected blood samples from many more grackles in Arizona, therefore we will redo the analyses from the Arizona population in the analyses involved in the current preregistration. In May 2020, we completed data collection for other variables at the Arizona field site: [flexibility and innovation](http://corinalogan.com/Preregistrations/g_flexmanip.html) [@logan2019flexmanip], and [exploration](http://corinalogan.com/Preregistrations/g_exploration.html) [@mccune2019exploration], and we will soon analyze this data, therefore it will also be considered secondary data. This preregistration was submitted in May 2020 to PCI Ecology for pre-study peer review. We received the reviews, and revised and resubmitted in Aug 2020, and it passed pre-study peer review in Oct 2020.~~

~~**Level of data blindness:** Logan and McCune collect the behavioral data (Q1) and therefore have seen this data for the Arizona population. Lukas has access to the Arizona data and has seen some of the summaries in presentations. Chen has not seen any data.~~

~~### B. PARTITIONING THE RESULTS~~

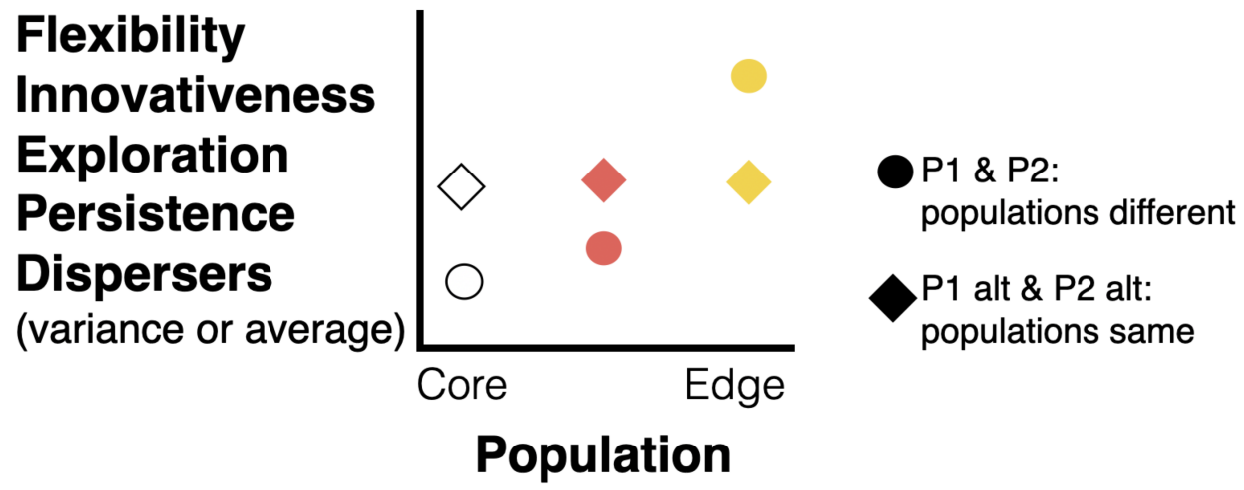
~~We may decide to present the results from different hypotheses in separate articles. We may also decide to test these hypotheses in additional species.~~

RESEARCH QUESTION: Are there differences in behavioral traits (flexibility, innovation, exploration, and persistence) between populations across the great-tailed grackle's geographic range? ~~(Fig. 1 & 2)~~.

Prediction 1: If behavior modifications are needed to adapt to new locations, then there is~~will be~~ a higher average and/or larger variance of at least some traits thought to be involved in range expansions (behavioral flexibility: speed at reversing a previously learned color preference based on it being associated with a food reward; innovativeness: number of options solved on a puzzle box; exploration: latency to approach/touch a novel object; and persistence: proportion of trials participated in with higher numbers indicating a more persistent individual) **in the grackles sampled from the more recently established population relative to the individuals sampled in**

the older populations (Table 1). Higher averages in behavioral traits indicate that each individual can exhibit more of that trait (e.g., they are more flexible/innovative/exploratory/persistent). Perhaps in newly established populations, individuals need to learn about and innovate new foraging techniques or find new food sources. Perhaps grackles require flexibility to visit these resources according to their temporal availability and the individual's food preferences. Perhaps solving such problems requires more exploration and persistence. Higher variances in behavioral traits indicate that there is a larger diversity of individuals in the population, which means that there is a higher chance that at least some individuals in the population could innovate foraging techniques and be more flexible, exploratory, and persistent, which could be learned by conspecifics and/or future generations. *This ~~would~~ supports the hypothesis* that changes in behavioral traits facilitate the great-tailed grackle's geographic range expansion.

~~**Prediction 1 alternative 1: Human-modified environments are suitable habitat for grackles** (e.g., Selander & Giller (1961), Johnson & Peer (2001), Wehtje (2003)), **and the amount of human-modified environments has increased and is increasing** (e.g., Liu et al. (2020)). If the original behaviors exhibited by this species happen to be suited to the uniformity of human-modified landscapes (e.g., urban, agricultural, etc. environments are modified in similar ways across Central and North America), then the averages and/or variances of these traits will be similar in the grackles sampled from populations across their range (Table 1). *This supports the hypothesis* that, because this species is closely associated with human-modified environments, which may be similar across the geographic range of this species, individuals in new areas may not need to learn very much about their new environment: they can eat familiar foods and access these foods in similar ways across their range (e.g., fast food restaurant chains likely make the same food and package it in the same packaging in Central and North America, outdoor cafes and garbage cans also look the same across their range). Alternatively, it is possible that 2.9 generations at the edge site is too long after their original establishment date to detect differences in the averages and/or variances (though evidence from experimental evolution suggests that, even after 30 generations there is no change in certain behaviors when comparing domestic guinea pigs with 30 generations of wild-caught captive guinea pigs Künzl et al. (2003), whereas artificial selection can induce changes in spatial ability in as little as two generations Kotrschal et al. (2013)). If the sampled individuals had already been living at this location for long enough (or for their whole lives) to have learned what they need about this particular environment (e.g., there may no longer be evidence of increased flexibility/innovativeness/exploration/persistence), there may be no reason to maintain population diversity in these traits to continue to learn about this environment. We will not be able to distinguish between these two alternatives within alternative 1 because populations closer to the northern edge of this species' range were too small for us to establish such a field site. Both of these alternatives assume that learning is costly (e.g., Mery & Kawecki, 2005), therefore individuals avoid it if they can. In the first case, individuals might not need to rely much on learning because they are attending to familiar cues across their range, therefore they only need to learn where in this new spacenew space space these cues are located. In the second case, individual learning that the founding individuals needed to rely on to move into this new space could have been lost due to potential pressure to reduce this investment as soon as possible after moving to a new location.~~



~~**Figure 1.** What is the role of behavior in a rapid range expansion? The great-tailed grackle study sites are indicated by the colored circles: edge (yellow; California), middle (red; Arizona), and core (white; Central America) and correspond with those in Figure 3.~~

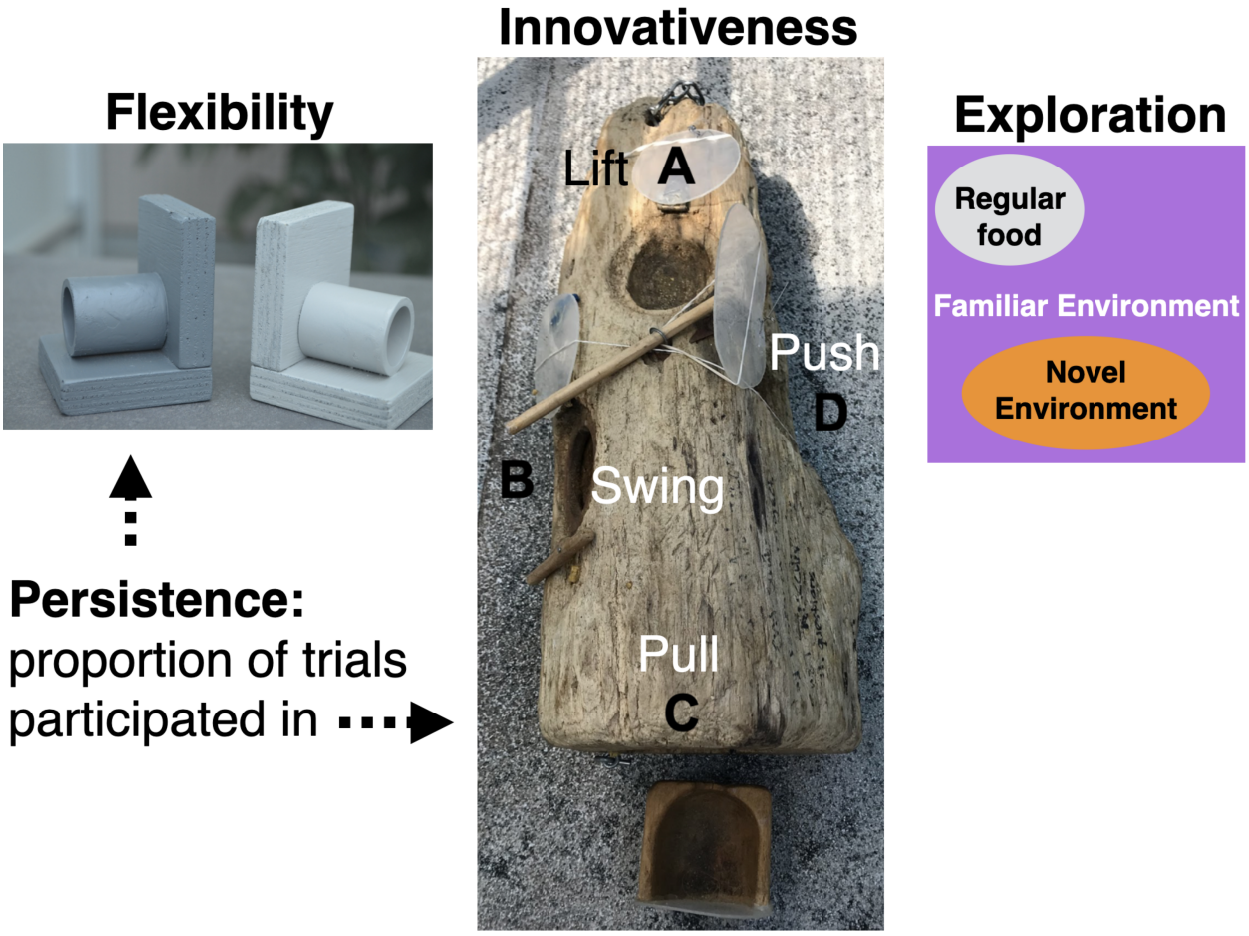


Figure 2. Experimental protocol. Great-tailed grackles from the ~~core, middle~~ older~~,~~ and newer~~edge~~ populations are~~will be~~ tested for their: (top left) flexibility (number of trials to reverse a previously learned color tube-food association); (middle) innovativeness (number of options [lift, swing, pull, push] solved to obtain food from within a multi-access log); (bottom left) persistence (proportion of trials participated in during flexibility and innovativeness tests); and (far right) exploration (latency to approach~~/touch~~ a novel environment~~object~~).

METHODS

~~Planned~~ Sample

~~**Q1 & Q2:**~~ Great-tailed grackles are caught in the wild in Woodland and in the Bufferlands of Sacramento, California~~and at a site to be determined in Central America~~. Some of our banded individuals were found in Woodland and the Bufferlands, therefore we consider this one population. We aim to bring adult grackles, rather than juveniles, temporarily into the aviaries for behavioral choice tests to avoid the potential confound of variation in cognitive development due to age, as well as potential variation in fine motor-skill development (e.g., holding/grasping objects; early-life experience plays a role in the development of both of these behaviors; e.g., @collias1964development, @rutz2016discovery) with variation in our target variables of interest. However, due to difficulties in trapping this species at this site, we also tested some juveniles. This should not pose a problem because we found that the two juveniles (Taco and Chilaquile) we tested

in the Tempe population did not perform differently from adults [@logan2023flexmanip; @blaisdell2021causal; @logan2021inhibition; @seitz2021touchscreentraining]. Adults awerewill be identified from their eye color, which changes from brown to yellow upon reaching adulthood (Johnson & and Peer, 2001). We apply colored leg bands in unique combinations for individual identification. Some individuals (33~20) are brought temporarily into aviaries for behavioral choice tests, and then are released back to the wild at their point of capture. We catch grackles with a variety of methods (e.g., walk-in traps and, mist nets, bow nets). Mist netssome of which decrease the likelihood of a selection bias for exploratory and bold individuals because grackles cannot see the traps (i.e., mist nets). Grackles are individually housed in an aviary (each 244 cm long by 122 cm wide by 213 cm tall) for 3 weeks to 6an allowable maximum of six 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 20 h per day), and various other food items (e.g., peanuts, bread) during testing (up to 4 h 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 occurs on a day off, they are tested on the fifth day instead).

While the above is our ideal plan was to conduct the same tests at an additional field site in Central America, due to restrictions around COVID-19 and also to issues with sexual abuse at the planned field site, it wasmay not be possible for us to accomplish thisall of our goals within our current funding period. We think it will be possible to collect data at one more site (which would be the second of three planned sites) and we will attempt to also include a third field site.

Sample size rationale

Q1 & Q2: We test as many great-tailed grackles as we can during the approximately one 2 years we spend at each of our fieldeach sites given that the birds are only brought into the aviaries during the non-breeding season (approximately September through AprilMarch). It is time intensive to conduct the aviary test battery (3 weeks2-6 months per bird at the Arizona field site), therefore we aim to meet approximate that the minimum sample size at each site will follow the minimum sample sizes in Supplementary Material Table SM12. Wewith the aim for an equal sex ratio of subjects to test (50%that half of the grackles tested at each site are of all female)s and achieved an overall 47% female (this percentage differsdiffered depending on the test). WHowever, we expected to be able to test 20 grackles per site. See the gxpopbehaviorhabitat_data_testhistory.csv data sheet at @logan2023xpopdata for a list of the order of experiments for each individual at the Woodland site, and g_flexmanip_data_AllGrackleExpOrder.csv at @logan2023flexmanipdata for the Tempe grackles.

Data collection stopping rule

We will stop collecting data on wild-caught great-tailed grackles in Q1 and Q2 (data for Q3 are collected from the literature) once we have completed one year at each of the California and Central America site and meet our minimum sample sizes (likely complete in summer 2022), which coincides with the period in which we currently have funding (until early 2023). If we are not able to collect data at a third site, we will attempt to collect more data during a second year at the second site (Woodland, CA).

Protocols and open materials

- Experimental protocols are online here.
- **Flexibility** protocol (from Logan et al., (2019)) using reversal learning with color tubes. Grackles are first habituated to a yellow tube and trained to search for hidden food. A light gray tube and a dark gray tube are placed on the table or floor: one color always contains a food reward (not visible by the bird) while the other color never contains a reward. The bird is allowed to choose one tube per trial. An individual is considered to have a preference if it choses the rewarded option at least 85% of the time (17/20 correct) in 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). We use a sliding window in 1-trial increments to calculate whether they passed after their first 20 trials. Once a bird learns to prefer one color, the contingency is reversed: food is

always in the other color and never in the previously rewarded color. The flexibility measure is how many trials it takes them to reverse their color preference using the same passing criterion.

- **Innovativeness** protocol (from Logan et al., ~~(~~2019~~)~~ and based on the experimental design by ~~AMI.~~ Auersperg et al., ~~(~~2011~~)~~) using a multi-access log. Grackles are first habituated to the log apparatus with all of the doors locked open and food inside each locus. After habituation, the log, which has four ways of accessing food (pull drawer, push door, lift door up, swing door out), is placed on the ground and grackles are allowed to attempt to solve or successfully solve one option per trial. Once a bird has successfully solved an option three times, it becomes non-functional (the door is locked open and there is no food at that locus). The experiment ends when all four loci become non-functional, if a bird does not come to the ground within 10 min in three consecutive test sessions, or if a bird does not obtain the food within 10 min (or 15 min if the bird was on the ground at 10 min) in three consecutive test sessions.
- **Persistence** is measured as the proportion of trials participated in during the flexibility and innovativeness experiments (after habituation, thus it is not confounded with neophobia). The higher the number, the more persistent they are. This measure indicates that those birds who do not participate as often are less persistent in terms of their persistence with engaging with the task. We generally offer a grackle the chance to participate in a trial for 5 min. If they do not participate within that time, we record -1 in the data sheet, the apparatus is removed, and the trial is re-attempted later.
- **Exploration** is measured as the latency to approach within 20 cm of a novel environment inside of their familiar aviary environment, averaged across Time 1 (on the individual's 8th day in the aviary) and Time 2 (1 week after Time 1). The bird's regular food is moved to one end of the aviary, away from the novel environment, and a motivation test precedes the session. The bird is then exposed to first a familiar environment (45 min) and then a novel environment (45 min). If an individual does not approach within 20 cm, it is given a latency of 2701 sec (45 min plus 1 sec).

Open data

The data and code are publicly available at the Knowledge Network for Biocomplexity's data repository (Logan ~~CJ~~ et al., 2023).

Randomization and counterbalancing

Experimental order: The order of experiments, reversal learning or multiaccess log, ~~was~~ counterbalanced across birds.

Reversal learning: The first rewarded color in reversal learning ~~was~~ counterbalanced across birds. The rewarded option ~~was~~ pseudorandomized for side (and the option on the left ~~was~~ always placed first). Pseudorandomization consists of alternating location for the first two trials of a session and then keeping the same color on the same side for at most two consecutive trials thereafter. A list of all 88 unique trial sequences for a 10-trial session, following the pseudorandomization rules, ~~was~~ generated in advance for experimenters to use during testing (e.g., a randomized trial sequence might look like: LRLLRRLRLR, where L and R refer to the location, left or right, of the rewarded tube). Randomized trial sequences ~~were~~ assigned randomly to any given 10-trial session using a random number generator (random.org) to generate a number from 1-88.

Analyses

We use~~d~~ **simulations** and design~~ed~~ customized **models** to determine what sample sizes allow us to detect differences between sites (Supplementary Material 2; see chapter 5.3 in Bolker (2008) for why simulations perform more powerful power analyses). We d~~id~~ not **exclude** any data, and data that ~~were~~ **missing** (e.g., if a bird participate~~d~~ in one of the two experiments) for an individual in a given experiment, then this individual ~~was~~ not included in that analysi~~e~~s. Analyses ~~were~~ conducted in R (current version 4.1.2; R Core Team, ~~(~~2017~~)~~) and Stan (version 2.18, Carpenter et al., ~~(~~2017~~)~~) using the following packages: psych (Revelle, 2017), irr (Gamer et al., 2012), rethinking (McElreath,

2020), rstan (Stan Development Team, 2020), knitr [@xie2018knitr; @xie2017dynamic; @xie2013knitr], dplyr [@dplyr], tidyr [@tidyr], cmdstanr [@cmdstanr], DHARMa [@hartig2019dharma], lme4 [@lme4; @bates2012lme4], and Rcpp (Eddelbuettel & François, 2011). Interobserver reliability scores indicate high agreement across coders for all dependent variables (see Supplementary Material 3 for details).

~~Interobserver reliability of dependent variables~~

~~To determine whether experimenters coded the dependent variables in a repeatable way, hypothesis-blind video coders were first trained in video coding the dependent variables (reversal learning and multiaccess log: whether the bird made the correct choice or not; exploration: latency to approach), requiring a Cohen's unweighted kappa (reversal and multiaccess categorical variables) or an intra-class correlation coefficient (ICC; exploration continuous variable) of 0.90 or above to pass training. This threshold indicated that the two coders (the experimenter and the video coder) agreed with each other to a high degree (kappa: Landis & Koch (1977); ICC: Hutcheon et al. (2010)). After passing training, the video coders coded 20% of the videos for each experiment and the kappa and ICC were calculated to determine how objective and repeatable scoring was for each variable, while noting that the experimenter has the advantage over the video coder because watching the videos is not as clear as watching the bird participate in the trial from the aisle of the aviaries. The unweighted kappa was used when analyzing a categorical variable where the distances between the numbers are meaningless (0=incorrect choice, 1=correct choice, -1=did not participate), and the ICC was used for continuous variables where distances are meaningful (e.g., if coders disagree by a difference of 2 s rather than 5 s, this is important to account for).~~

~~Interobserver reliability training~~

~~To pass **interobserver reliability (IOR) training**, video coders needed an ICC or Cohen's unweighted kappa score of 0.90 or greater to ensure the instructions were clear and that there was a high degree of agreement across coders. Video coders, Alexis Breen and Vincent Kiepsch, passed interobserver reliability training for exploration in a previous article (McCune KB et al., 2019) where their training results can be found.~~

~~***Lea Gihlein*** (compared with experimenter's live coding):~~

- ~~Reversal learning: correct choice unweighted Cohen's Kappa=1.00 (confidence boundaries=1.00-1.00, n=21 data points)~~
- ~~Multiaccess box: correct choice unweighted Cohen's Kappa=1.00 (confidence boundaries=1.00-1.00, n=29 data points)~~
- ~~Multiaccess box: correct choice unweighted Cohen's Kappa=1.00 (confidence boundaries=1.00-1.00, n=29 data points)~~

~~Interobserver reliability~~

~~Interobserver reliability scores (minimum 20% of the videos) were as follows:~~

~~***Lea Gihlein*** (compared with experimenter's live coding):~~

- ~~Reversal learning (5/19 birds): correct choice unweighted Cohen's Kappa=1.00 (confidence boundaries=0.99-1.00, n=707 data points)~~
- ~~Multiaccess box (5/23 birds): correct choice unweighted Cohen's Kappa=0.92 (confidence boundaries=0.81-1.00, n=63 data points)~~
- ~~Multiaccess box (5/23 birds): locus solved unweighted Cohen's Kappa=1.00 (confidence boundaries=1.00-1.00, n=48 data points)~~

~~***Vincent Kiepsch*** (compared with Breen's video coding):~~

- ~~Exploration (5/34 birds): latency to land on the ground unweighted Cohen's Kappa=0.998 (confidence boundaries=0.997-0.999, n=32 data points)~~

~~Code~~

~~Hypothesis-specific mathematical model~~

~~Following procedures in McElreath (2016), we constructed a **hypothesis-appropriate mathematical model** for each of the response variables that examines differences in the response variable between sites (each site represents a grackle population). These models take the form of:~~

~~y ~ α[site]~~

~~y is the response variable (flexibility, innovation, exploration, or persistence). There iswill be one intercept, α, per site and we will estimate the site's average and standard deviation of the response variable.~~

~~We formulated these models in a Bayesian framework. We determined the priors for each model by performing prior predictive simulations based on ranges of values from the literature to check that the models are covering the likely range of results.~~

~~We will then perform pairwise contrasts to determine at what point we~~ can~~will be able to detect differences between sites by manipulating sample size, and α means~~, and standard deviations. ~~Before running the simulations, we decided that a model would detect an effect if 89% of the difference between two sites is on the same side of zero (following McElreath (2016)). We are using a Bayesian approach, therefore comparisons are based on samples from the posterior distribution. We will draw 10,000 samples from the posterior distribution, where each sample will hasve an estimated mean for each population. For the first contrast, within each sample, we subtract the estimated mean of the edge population from the estimated mean of the core population. For the second contrast, we subtract the estimated mean of the edge population from the estimated mean of the middle population. For the third contrast, we subtract the estimated mean of the middle population from the estimated mean of the core population. We will now have samples of differences between all of the pairs of sites, which we can use to assess whether any site is systematically larger or smaller than the others. We will determine whether this is the case by estimating what percentage of each sample of differences is either larger or smaller than zero. For the first contrast, if 89% of the differences are larger than zero, then the core population has a larger mean. If 89% of the differences are smaller than zero, then the edge population has a larger mean.~~

Flexibility analyses

Model and simulation

We modifi~~ed~~ the reversal learning Bayesian model in Blaisdell et al. (2021) to simulate and analyze population differences in reversal learning, and calculate our ability to detect differences between populations. The model accounts for every choice made in the reversal learning experiment and updates the probability of choosing either option after the choice i~~wa~~s made depending on whether that choice contain~~sed~~ a food reward or not. It does this by updating three main components for each choice: an attraction score, a learning rate (ϕ), and a rate of deviating from learned attractions (λ).

~~**Equation 1 (attraction and ϕ):** $A_{i,j,t+1} = (1-\phi_j)A_{i,j,t} + \phi_j \pi_{i,j,t}$~~

~~Equation 1 "tells us how attractions to different behavioral options $A_{i,j,t+1}$ (i.e., how preferable option i is to the bird j at time t+1) change over time as a function of previous attractions $A_{i,j,t}$ and recently experienced payoffs $\pi_{i,j,t}$ (i.e., whether they received a reward in a given trial or not). Attraction scores thus reflect the accumulated learning history up to this point. The (bird-specific) parameter ϕ_j describes the weight of recent experience. The higher the value of ϕ_j, the faster the bird updates their attraction. It thus can be interpreted as the *learning or updating rate of an individual*. A value of ϕj=0.04, for example, means that receiving a single reward for one of the two options will shift preferences by 0.02 from initial 0.5-0.5 attractions, a value of ϕj=0.06 will shift preferences by 0.03 and so on" (Blaisdell et al., 2021).~~

~~**Equation 2 (λ):**~~

~~$P(i)_{t+1} = \exp(\lambda_j A_{i,j,t}) / \sum_{m=1}^{2} \exp(\lambda_j A_{m,j,t})$.~~

P(i)t+1=exp(λjAi,j,t)∑m=12exp(λjAm,j,t).

Equation 2 "expresses the probability an individual *jj* chooses option *ii* in the next round, *t*+1

t+1, based on the latent attractions. The parameter λ_jλj represents the *rate of deviating from learned attractions* of an individual (also called inverse temperature). It controls how sensitive choices are to differences in attraction scores. As λ_jλj gets larger, choices become more deterministic, as it gets smaller, choices become more exploratory (random choice if λ_j=0). For instance, if an individual has a 0.6-0.4 preference for option A, a value of λ_j=3 means they choose A 65% of the time, a value of λ_j=10 means they choose A 88% of the time and a value of λj=0.5 means they choose A only 53% of the time" (Blaisdell et al., 2021).

As in Blaisdell et al. (2021), we, too, used previously published data on reversal learning of color tube preferences in great-tailed grackles in Santa Barbara, California (Logan, 2016b) to inform the model modifications. We modifyied the Blaisdell et al. (2021) model in twoa few ways: 1) we set the initial attraction score assigned to option 1 and option 2 (the light gray and dark gray tubes) to 0.1 rather than 0.0. This change assumes that there would be some inclination (rather than no inclination) for the bird to approach the tubes when they are first presented because they arehad been previously trained to expect food in tubes. This also allows the attraction score to decrease when a non-rewarded choice is made near the beginning of the experiment. With the previous initial attraction scores set to zero, a bird would be expected to choose the rewarded option in 100% of the trials after the first time it chose that option (attraction cannot be lower than zero, and choice is shaped by the ratio of the two attractions so that when one option is zero and the other is larger than zero, the ratio will be 100% for the rewarded option). 2) We changed the updating so that an individual would only changes the attraction toward the option they chose in that trial (either decreasing their attraction toward the unrewarded option or increasing their attraction toward the rewarded option). Previously, both attractions were updated after every trial, assuming that individuals understand that the experiment is set upsetup such that one option is always rewarded. For our birds, we instead assumed that individuals will focus on their direct experience rather than making abstract assumptions about the test. Our modification resulted in needing a higher ϕ to have the same learning rate as a model where both attraction scores are updated after every trial. This change also appearsed to better reflect the performance of the Santa Barbara grackles, because they had higher ϕ values, which, in turn, meanst lower λ values to reflect the performance during their initial learning. These lower λ values better reflect the birds' behavior during the first reversal trials: a large λ value means that birds continue to choose the now unrewarded option almost 100% of the time, whereas the lower λ values mean that birds start to explore the rewarded option relatively soon after the switch of the rewarded option.

We first reanalyzed the Santa Barbara grackle data to obtain the phi and lambda values with this revised model, which informsed our expectations of what a site's mean and variance might be. Then we ruan simulations, where we determined that we wanted to make the previously mentioned modificationsmodificatio9s to the stan (Team et al., 2019) model [in R, current version 4.1.2; R Core Team (2017)]. This model iwsill be used to analyze the actual data after it is collected.

Code

Power analyses: We also used the simulations to estimate our ability to detect differences in ϕϕ and λλ between sites based on extracting samples from the posterior distribution. We ruan two different sets of simulations: the first set of simulations recreated choices for 20 birds per population across initial learning and reversal trials; the second set of simulations we first sampled between 9 and 24 birds from populations with pre-specified ϕϕ and λλ means to determine the minimum sample size required to detect whether two populations are different. This set of simulations shows how different site sample sizes change detection levels: once a sample size of 15 is reached, there are only minimal differences in detection abilities compared to larger sample sizes (Figure 446). The second set of simulations recreatesrecreatsed choices for 20 birds per population across initial learning and reversal trials from which we estimated their ϕϕ and λλ. We chose to simulate 20 birds per population because this number is above the threshold we detected in the first set of simulations and it appearsed a feasible sample size. We expected that the noise in the probabilistic choices of individuals might reduce the differences that can be detected compared to the first simulation where ϕϕ and λλ awere assumed to be exactly known for each individual. This second e first set of

simulations showed that we have a very high chance of detecting that two sites are different from each other if the difference in their ϕϕ is 0.01 or greater and/or if the difference in their λλ is 3 or greater, based on data from 20 simulated individuals per site (Figure 35). The second set of simulations shows how different site sample sizes change detection levels: once a sample size of 15 is reached, there are only minimal differences in detection abilities compared to larger sample sizes (Figure 46). It appears that there is more variability in the λλ estimates for each bird based on their choices, meaning that with the learning model, which estimates λλ from the choices, the differences between sites have to be larger (than if we were able to infer lambda directly) to be reliably detected. Given that we have to infer ϕϕ and λλ from the choicesThe second set of simulations assumed that we could infer λλ directly, however it is not possible to directly measure λλ. Therefore, the power curves in Figure 35 are more reliable than those in Figure 46.

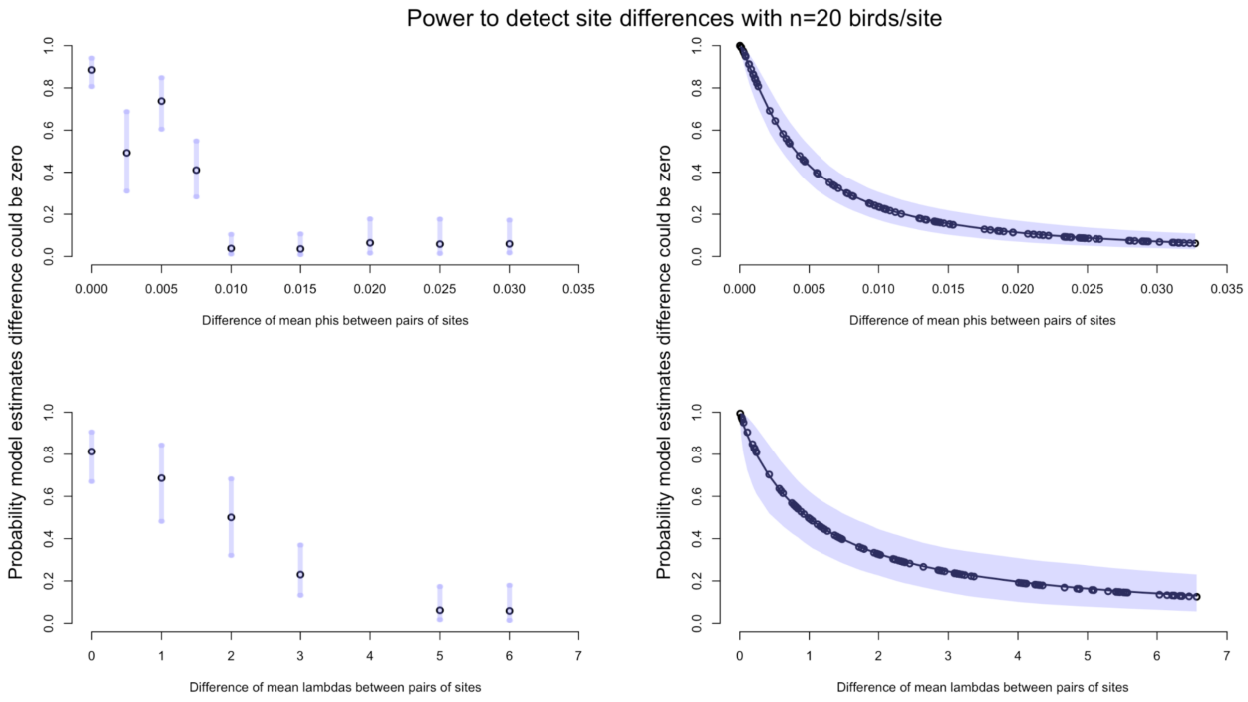


Figure 35. How small of a site difference in phi and lambda can we detect? The probability that the model estimates that the difference shown on the x axis is zero, meaning that the model assumes that it is possible that these two estimates come from a population with the same phi or lambda. Each point is the mean phi or mean lambda from one site minus the mean phi or mean lambda from another site (calculated from 20 individuals per site) for all pairwise comparisons for all 32 simulated sites (for a total of 496 pairwise comparisons). Left panels: error bars=89% compatiability intervals. Right panels: shaded areas=97% prediction intervals.

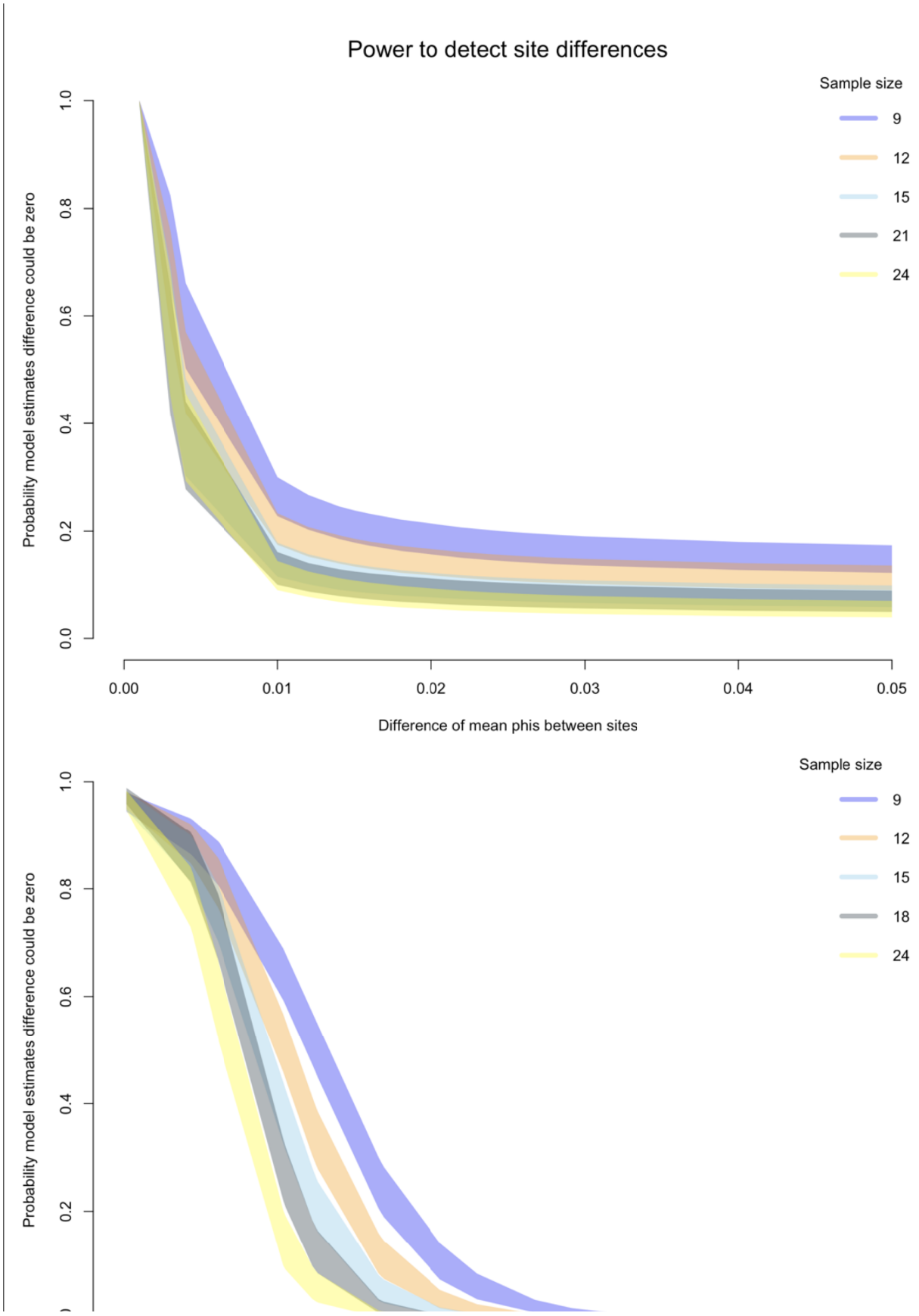

Power to detect site differences
Sample size
9
12
15
21
24
Probability model estimates difference could be zero
0.0
0.2
0.4
0.6
0.8
1.0
0.00
0.01
0.02
0.03
0.04
0.05
Difference of mean phis between sites
Sample size
9
12
15
18
24
Probability model estimates difference could be zero
0.2
0.4
0.6
0.8
1.0

Figure 46. How do detection differences vary according to sample size differences? The probability that the model estimates that the difference shown on the x axis is zero, meaning that the model assumes that it is possible that these two estimates come from a population with the same phi or lambda. The x-axis is the mean phi or mean lambda from one site minus the mean phi or mean lambda from another site for all pairwise comparisons for all 14 sites (for a total of 91 pairwise comparisons). Each shaded region is the 97% prediction interval for that particular sample size.

Code

Innovation analysis

Model and simulation

Expected values for the number of options solved on the multiaccess log awere set to 0-4 (out of 4 options maximum) because this apparatus had been used on two species of jays who exhibited individual variation in the number of loci solved between 0-4 (California scrub-jays (Aphelocoma californica) and Mexican jays (Aphelocoma wollweberi): McCune, (2018); McCune et al., (2019)).

locisolved ~ Binomial(4, p) *[likelihood]*

logit(p) ~ α[site] *[model]*

locisolved is the number of loci solved on the multiaccess box, 4 is the total number of loci on the multiaccess box, p is the probability of solving any one locus across the whole experiment, α is the intercept, and each site gets its own intercept. After running simulations, we identified the following distribution to be the most likely priors for our expected data:

αα ~ Normal(0,1) *[α prior]*

We used a normal distribution for α because it is a sum (see Figure 10.6 in McElreath (2016)) and a logit link to ensure the values are between 0 and 1. We set the mean to 0 on a logit scale, which means an individual solves 2 loci on average on the actual scale at a probability of 0.5.

Note that two grackles, Kau and Galandra, were accidentally able to pull 2 and 1, respectively, locus doors open during habituation to the multi-access box. Because habituation was not observed by an experimenter, the birds had the possibility to learn how these doors worked. Therefore, these doors were locked open and non-functional throughout their entire experiment. We accommodated for this in the model by replacing the 4 (as in 4 possible loci were available to solve) with a column of data that listed the maximum possible loci available to each bird.

Code

We then ruan the **mathematical model** and performed pairwise contrasts and determined that we arewill be able to detect differences between sites with a sample size of 15 at each site if the average number of loci solved differs by 1.2 loci or more and the standard deviation is generally a maximum of 0.9 at each site (see Supplementary Material 3 for detailsTable 3). For a sample size of 20 at each site, we arewill be able to detect site differences if the average number of loci solved differs by 0.7 of a locus or more and the standard deviation is generally a maximum of 1 at each site (Table SM33). Note: the Arizona sample size is 11 for the multiaccess log and 17 on a similar multiaccess box.

Table 3. Simulation outputs from **varying sample size (n), and** αα **means and standard deviations**. We calculate pairwise contrasts between the estimated means from the posterior distribution: if for a large sample the difference is both positive and negative and crosses zero (yes), then we are not able to detect differences between the two sites. If the differences between the means are all on one side of zero for 89% of the posterior samples (no), then we are able to detect differences between the two sites. We chose the 89% interval based on (McElreath, 2016). Note that for latency, there is no mu_sd, but rather one phi that is the same for all sites. mu=average, sd=standard deviation. The numbers 1-3 in the column titles refer to sites 1-3 as do S1-3 (the simulations were run on a total of three sites because we originally planned to collect data at two to three sites), mu=average, sd=standard deviation. Loci solved is the innovativeness measure, latency is the exploration measure, and trials participated in is the persistence measure.

Code

Response variable	n	mu1	mu2	mu3	mu1_sd	mu2_sd	mu3_sd	Difference crosses zero? S1-S2	Difference crosses zero? S1-S3	Difference crosses zero? S2-S3	Notes
loci solved	60	1.90	2.10	3.60	0.50	0.50	0.50	No	No	No	
loci solved	60	1.90	2.10	2.20	0.50	0.50	0.50	Yes	Yes	Yes	
loci solved	60	1.90	2.10	2.30	0.50	0.50	0.50	Yes	Yes	Yes	
loci solved	60	1.90	2.10	3.50	0.50	0.50	0.50	No	No	No	
loci solved	60	1.90	2.10	3.00	0.50	0.50	0.50	Yes	No	No	
loci solved	60	1.90	2.10	2.80	0.50	0.50	0.50	Yes	No	Yes	
loci solved	60	1.80	2.10	3.00	0.50	0.50	0.50	Yes	Yes	No	
loci solved	60	2.00	2.50	3.00	0.50	0.50	0.50	No	Yes	No	
loci solved	60	2.00	2.50	3.10	0.50	0.50	0.50	No	No	Yes	

loci solved	60	1.90	2.50	3.20	0.50	0.50	0.50	Yes	No	No
loci solved	60	1.80	2.50	3.30	0.50	0.50	0.50	No	No	Yes
loci solved	60	1.70	2.50	3.40	0.50	0.50	0.50	No	No	No
loci solved	60	1.70	2.50	3.40	1.00	1.00	1.00	No	No	No
loci solved	60	1.70	2.50	3.40	1.50	1.50	1.50	Yes	No	No
loci solved	60	1.70	2.50	3.40	1.30	1.30	1.30	Yes	Yes	Yes
loci solved	60	1.00	2.00	3.00	0.50	0.50	0.50	No	No	Yes
loci solved	60	1.00	2.00	3.00	0.50	0.50	0.50	No	No	No
loci solved	60	1.00	2.00	3.00	0.30	0.40	0.50	No	No	No
loci solved	60	1.00	2.00	3.00	0.60	0.70	0.50	No	No	No
loci solved	60	1.00	2.00	3.00	0.70	0.70	0.70	No	No	No

loci solved	60	1.00	2.00	3.00	0.90	0.90	0.90	No	No	No
loci solved	60	1.00	2.00	3.00	1.00	1.00	1.00	No	No	No
loci solved	60	1.00	2.00	3.00	1.50	1.50	1.50	Yes	No	No
loci solved	60	1.00	2.00	3.00	1.30	1.50	1.50	Yes	No	No
loci solved	60	1.00	2.00	3.00	1.10	1.50	1.50	No	No	No
loci solved	60	1.00	2.00	3.00	1.20	1.50	1.50	No	No	No
loci solved	45	1.00	2.00	3.00	0.50	0.50	0.50	Yes	No	No
loci solved	45	0.90	2.00	3.10	0.50	0.50	0.50	No	No	Yes
loci solved	45	0.80	2.00	3.20	0.50	0.50	0.50	No	No	No
loci solved	45	0.80	2.00	3.20	1.00	1.00	1.00	Yes	No	No
loci solved	45	0.80	2.00	3.20	0.90	0.90	0.90	No	No	No

latency	45	5.70	6.90	7.60	1000.00	1000.00	1000.00	No	No	No
latency	45	5.80	6.90	7.50	1000.00	1000.00	1000.00	No	No	No
latency	45	6.00	6.90	7.20	1000.00	1000.00	1000.00	No	No	Yes
latency	45	6.00	6.90	7.30	1000.00	1000.00	1000.00	No	No	Yes
latency	45	6.00	6.90	7.40	1000.00	1000.00	1000.00	No	No	Yes
latency	45	6.00	6.90	7.50	1000.00	1000.00	1000.00	Yes	No	No
latency	45	5.90	6.90	7.50	1000.00	1000.00	1000.00	No	No	Yes
latency	45	5.90	6.90	7.60	1000.00	1000.00	1000.00	No	No	No
latency	45	5.90	6.90	7.60	1000.00	1000.00	1000.00	No	No	No
latency	45	5.90	6.90	7.60	1000.00	1000.00	1000.00	No	No	No
latency	45	4.60	6.30	7.10	1000.00	1000.00	1000.00	No	No	Yes

latency	45	4.60	6.30	7.20	1000.00	1000.00	1000.00	No	No	No
latency	45	4.60	6.30	7.20	1000.00	1000.00	1000.00	No	No	Yes
latency	45	4.60	6.30	7.20	1000.00	1000.00	1000.00	No	No	Yes
latency	45	4.60	6.30	7.30	1000.00	1000.00	1000.00	No	No	No
latency	45	4.60	6.30	7.30	1000.00	1000.00	1000.00	No	No	No
latency	45	4.60	6.30	7.30	1000.00	1000.00	1000.00	No	No	No
latency	60	5.70	6.90	7.60	1000.00	1000.00	1000.00	No	No	Yes
latency	60	5.70	6.90	7.60	1000.00	1000.00	1000.00	No	No	Yes
latency	60	5.70	6.90	7.70	1000.00	1000.00	1000.00	No	No	No
latency	60	5.90	6.90	7.60	1000.00	1000.00	1000.00	No	No	Yes
latency	60	5.90	6.90	7.60	1000.00	1000.00	1000.00	No	No	Yes

latency	60	4.60	6.20	7.10	1000.00	1000.00	1000.00	Yes	No	No
latency	60	4.60	6.20	7.10	1000.00	1000.00	1000.00	Yes	No	No
latency	60	4.60	6.40	7.10	1000.00	1000.00	1000.00	No	No	No
latency	60	4.60	6.30	7.10	1000.00	1000.00	1000.00	No	No	Yes
latency	60	4.60	6.30	7.20	1000.00	1000.00	1000.00	No	No	Yes
latency	60	4.60	6.30	7.30	1000.00	1000.00	1000.00	No	No	Yes
latency	60	4.60	6.30	7.30	1000.00	1000.00	1000.00	No	No	No
latency	60	4.60	6.30	7.30	1000.00	1000.00	1000.00	No	No	No
latency	60	4.60	6.30	7.30	1000.00	1000.00	1000.00	No	No	Yes
latency	60	4.60	6.30	7.40	1000.00	1000.00	1000.00	No	No	No
latency	60	4.60	6.30	7.40	1000.00	1000.00	1000.00	No	No	No

latency	60	4.60	6.30	7.40	1000.00	1000.00	1000.00	No	No	No
latency	60	4.70	6.30	7.40	1000.00	1000.00	1000.00	Yes	No	No
trials participated	45	0.65	0.75	0.90	0.25	0.25	0.25	No	No	No
trials participated	45	0.70	0.75	0.82	0.25	0.25	0.25	Yes	No	No
trials participated	45	0.67	0.75	0.85	0.25	0.25	0.25	No	Yes	No
trials participated	45	0.67	0.75	0.90	0.25	0.25	0.25	No	No	No
trials participated	45	0.67	0.75	0.90	0.25	0.25	0.25	No	No	No
trials participated	45	0.67	0.75	0.90	0.25	0.25	0.25	No	No	No
trials participated	45	0.67	0.75	0.90	0.35	0.35	0.35	Yes	No	No

trials participated	45	0.67	0.75	0.90	0.30	0.30	0.30	Yes	No	No
trials participated	60	0.67	0.75	0.90	0.30	0.30	0.30	No	No	No
trials participated	60	0.70	0.75	0.81	0.30	0.30	0.30	No	Yes	No
trials participated	60	0.70	0.75	0.81	0.25	0.25	0.25	No	Yes	No
trials participated	60	0.68	0.75	0.83	0.25	0.25	0.25	Yes	No	No
trials participated	60	0.67	0.75	0.83	0.25	0.25	0.25	No	No	Yes
trials participated	60	0.67	0.75	0.83	0.25	0.25	0.25	No	No	No
trials participated	60	0.67	0.75	0.83	0.25	0.25	0.25	No	No	No

~~trials-participated~~	~~60~~	~~0.67~~	~~0.75~~	~~0.83~~	~~0.25~~	~~0.25~~	~~0.25~~	~~No~~	~~No~~	~~No~~	
~~trials-participated~~	~~60~~	~~0.67~~	~~0.75~~	~~0.84~~	~~0.25~~	~~0.25~~	~~0.25~~	~~Yes~~	~~No~~	~~Yes~~	~~Model breaks if I increase the SD to 0.3~~

~~Because the mean and the variance are linked in the binomial distribution, and because the variance simulations in the flexibility analysis showed that we arewill not be able to robustly detect differences in variance between sites, we will plot the variance in the number of loci solved between sites to determine whether the edge population has a wider or narrower spread than the other two populations.~~

Code

Exploration analysis

Model and simulation

We model~~ed~~ the average latency to approach a novel environment~~object~~ and compare~~d~~ these between sites. We simulate~~d~~ data and set the model as follows:

latency ~ gamma-Poisson(λ_i,ϕ) *[likelihood]*

log(λ_i) ~ α[site] *[the model]*

latency is the average latency to approach a novel environment~~object~~, λ_i is the rate (probability of approaching the novel environment~~object~~ in each second) per bird (and we take the log of it to make sure it is always positive; birds with a higher rate have a smaller latency), ϕ is the dispersion of the rates across birds, and α is the intercept for the rate per site.

Expected values for the latency to approach a novel environment~~object~~ range from 0-2700 sec, which encompasses the time period during which they are exposed to the novel environment~~object~~ (sessions last up to 45 min). However, we do not provide an upper limit for the model because those birds that do not approach within 2700 sec would eventually have had to approach the novel environment~~object~~ to access their food (it is just that sessions do~~id~~ not run that long). After running simulations, we identify~~ied~~ the following distribution and priors to be the most likely for our expected data:

ϕ ~ 1/(Exponential(1)) *[ϕ prior]*

α ~ Normal(1350,500) *[α prior]*

We use~~d~~ a gamma-Poisson distribution for latency because it constrains the values to be positive. For ϕ, we use~~d~~ an exponential distribution because it is standard for this parameter~~paramter~~. We use~~d~~ a normal distribution for αα because it is a sum with a large mean (see Figure 10.6 in McElreath (2016)). We estimate that the grackles might approach the novel environment~~object~~ at any time in the session, therefore we ho~~e~~ld the αα mean of 1350 sec in mind as we conduct~~ed~~ the

modeling. We set the αα standard deviation to 500 because this puts the range of seconds for the distribution in the possible range.

Code

We then ruan the **mathematical model** and performed pairwise contrasts and determined that we arewill be able to detect differences between sites with a potential sample size of 15 at each site or 20 at each site if the average latency to approach the novel environmentobject differs by at least 450 sec betweenat each sites (Table 3). We keept the shape of the curve (which can be thought of as similar to a standard deviation or the variance) the same across sites because we do not think this assumption will change across populations (i.e., there will be lots of variation at each site with some individuals approaching almost immediately, others in the middle of the session, and others near the end).

Because the mean and the variance are linked in the gamma-Poisson distribution, and because the variance simulations in the flexibility analysis showed that we will not be able to robustly detect differences in variance between sites, we will plot the variance in the latency to approach the novel environmentobject between sites to determine whether the edge population has a wider or narrower spread than the other two populations.

Results (using our actual data)

We will analyze our data using the above model once all of the data have been collected.

Persistence analysis

Model and simulation

Expected values for the number of trials not participated in canould range from 0-125. The (likely maxima for reversal learning is: 300 trials reversal learning [approximate maxima based on data from Santa Barbara [@logan2016flexibilityproblem] and Tempe grackles [@logan2023flexmanip] where,: on average, individuals participate in 70 trials in theinitial discrimination, a maximum of 130 trials in the reversal, and up to ~100 non-participation trials across the initial discrimination and reversaltwo stages],. On the multiaccess log, grackles participated in a maximum ofplus 50 trials multiaccess log and there were up to [~25 non-participation trials with the log]). The estimated maximum number of non-participation trials is based on what might be expected from an individual who does not participate very often. After running simulations, we identifyied the following distribution and priors most likely for our expected data:

participated ~ Binomial(totaltrials, p) *[likelihood]*

logit(p) ~ α[site] *[model]*

participated indicates whether the bird participated or not in a given trial, total trials is the total number of trials offered to the individual (those participated in plus those not participated in), p is the probability of participating in a trial, αα is the intercept, and each site gets its own intercept. We used a logit link to constrain the output to between 0 and 1. After running simulations, we identifyied the following distribution and priors most likely for our expected data:

α ~ Normal(0,0.5) *[α prior]*

We used a normal distribution for αα because it is a sum (see Figure 10.6 in McElreath, (2016)). We set the mean to 0 (on a logit scale, which is a probability of 0.5 that a bird will participate in every other trial on average on the actual scale).

Code

We then ruan the **mathematical model** and performed pairwise contrasts and determined that we arewill be able to detect differences between sites with a potential sample size of 15 per site or 20 per site if the average proportion of trials participated in differs by at least 0.08 and the standard deviation is generally a maximum of 0.25 at each site (Table SM33).

Because the mean and the variance are linked in the binomial distribution, and because the variance simulations in the flexibility analysis showed that we arewill not be able to robustly detect differences

~~in variance between sites, we will plot the variance in the proportion of trials participated in between sites to determine whether the edge population has a wider or narrower spread than the other two populations.~~

Repeatability of exploration 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 (bird 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.

Post-study choices made since receiving in principle recommendation

In the preregistration, we said that for the exploration measure we would use the "Latency to approach within 20 cm of an object (novel or familiar, that does not contain food) in a familiar environment (that contains maintenance diet away from the object) - OR - closest approach distance to the object (choose the variable with the most data for the analysis)." We had data for both exploration measures and we used the latency measure because this was the variable that our preregistered analysis was designed for.

In the peer review history of the preregistration, we said that we would use whichever exploration test was repeatable with the Tempe grackles (novel object and/or novel environment) (round 1, response 16, https://ecology.peercommunityin.org/articles/rec?id=98). The methods for both novel stimuli were exactly the same and there was little variation in whether, or for how long, individuals went into the novel environment (i.e., most individuals did not go in the novel environment). However, the Tempe grackles responded differently to the novel environment and novel object, therefore they did not perceive the stimuli as the same. From the Tempe grackle data, we found that responses were only repeatable for the novel environment test [@mccune2019exploration]. Therefore, we conducted this assay (and not the novel object assay) with the Woodland grackles and compared the two populations on this one assay.

For the repeatability of persistence, the preregistered model had Test (reversal or multiaccess box) as the explanatory variable and ID as the random variable. However, we believe we made an error in choosing the explanatory variable because we are interested in whether the trait is repeatable across populations regardless of test. Therefore, we replaced Test with Population in the model. In addition, we realized that our measure of persistence (proportion of trials participated in) is not appropriate for a Poisson model, as preregistered. Consequently, we use a likelihood ratio test to compare a mixed model to a model without the ID random effect, and the function rpt from the package: rptR [@stoffel2017rpt] to estimate the variance in the dependent variable attributable to consistent differences among individuals across the two tests. We previously found that this method produces the same repeatability results as the MCMCglmm method using a Gaussian distribution [@mccune2023flexmanip].

The exploration data for the repeatability calculation were heteroscedastic and overdispersed. Additionally, 53% of the data were at the ceiling value (i.e., the bird did not approach the novel environment). Consequently, the model that best fit the data and was appropriate for the repeatability analysis was a binomial model, where the response was 0 (the grackle never approached the novel environment during exploration trials) or 1 (the grackle approached the novel environment).

RESULTS

Flexibility

There are no strong site differences for either component of reversal learning: phi or lambda (Figure 3). However, phi differs by only 0.0005 (Woodland=0.0306, Tempe=0.0301) and lambda by 0.26 (Woodland=4.78, Tempe=4.52), and the compatibility intervals for the estimated differences for both parameters cross zero (Table 2). With our sample size, we only have the power to reliably detect differences between the populations if they are larger than 0.01 for phi and 3 for lambda (based on our power analysis in Supplementary Material 2, summarized in Table SM1). Accordingly, we cannot exclude that the two populations are different, however we can estimate the range for how small the difference can be. Based on the estimated 89% compatibility intervals for phi and lambda in Table 2, the two populations are unlikely to differ by more than 0.01 for phi and 1.4 for lambda. Plotting the values (Figure 3) suggests no differences in the variances because similar minimum and maximum values are observed in both populations.

Table 2. Contrasts (indicated by "diff") between populations for the flexibility measure of reversal learning: phi and lambda.

	Mean	Standard deviation	Lower 89 percentile compatibility interval (5.5%)	Upper 89 percentile compatibility interval (94.5%)
Woodland Phi	0.03	0.01	0.02	0.05
Woodland Lambda	5.84	5.96	1.79	12.00
Tempe Phi	0.03	0.01	0.02	0.04
Tempe Lambda	5.51	3.93	1.43	11.40
diff_Phi	0.00	0.01	-0.01	0.01
diff_Lambda	0.26	0.68	-0.73	1.40

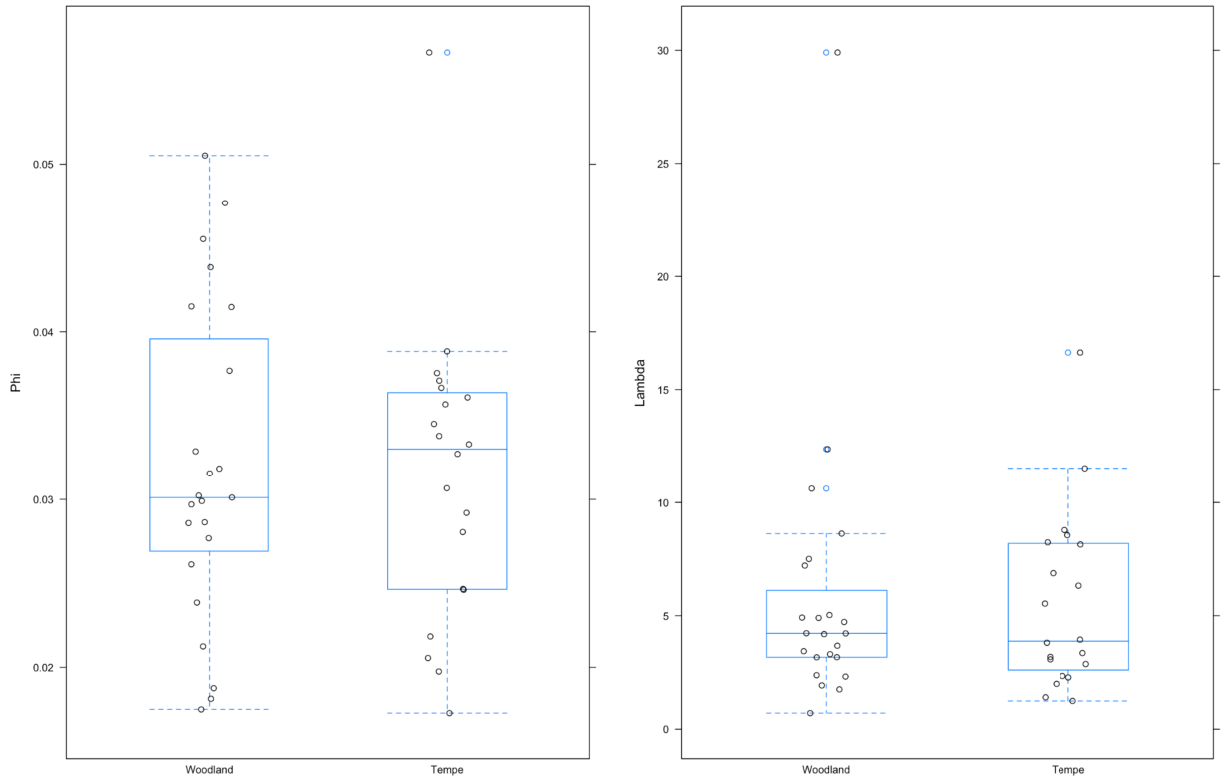


Figure 3. Measures of flexibility from the reversal learning experiment: phi and lambda per individual in each population. The boxplots show the minimum, maximum, lower and upper quartiles, and median values. The blue circles are outliers associated with the boxplots. The black circles are the raw data from each individual.

Innovation

Individuals in the more recent population, Woodland, California, are more innovative than individuals in the older population in Tempe, Arizona (Figure 4). Woodland grackles solve a higher proportion of loci on the multiaccess box as indicated by the contrast that showed that the compatibility interval did not cross zero (diff_12 in Table 3). Plotting the values (Figure 4) suggests no clear differences in the variances between the two populations because some individuals in both populations solved zero and some solved all four loci.

Table 3. Contrasts (indicated by "diff") between populations for the innovation measure: the proportion of loci solved on the multiaccess box.

	Mean	Standard deviation	Lower 89 percentile compatibility interval (5.5%)	Upper 89 percentile compatibility interval (94.5%)
Woodland	0.76	0.04	0.69	0.83
Tempe	0.50	0.06	0.41	0.60
diff_12	0.26	0.07	0.14	0.37

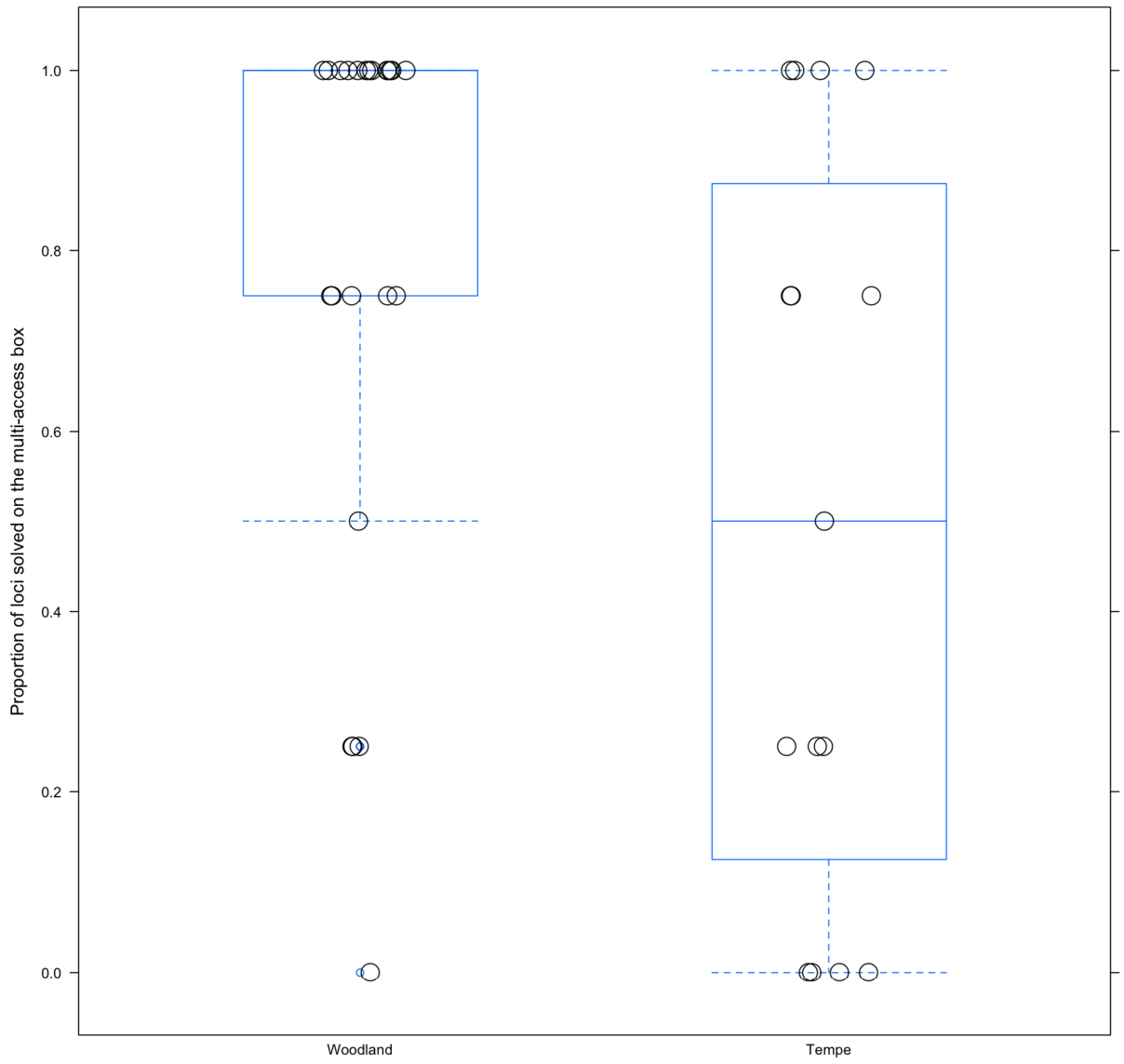


Figure 4. Number of loci solved on the multiaccess box in the innovativeness test per individual at each site (n=21 birds in Woodland, n=15 birds in Tempe). The boxplots show the minimum, maximum, lower and upper quartiles, and median values. The blue circles are outliers associated with the boxplots. The black circles are the raw data from each individual.

Exploration

Individuals in the older population, Tempe, Arizona, are more exploratory than individuals in the more recent population in Woodland, California (Figure 5). Tempe grackles are faster (have lower latencies) to approach a novel environment as indicated by the contrast that shows that the compatibility interval does not cross zero (diff_12 in Table 4). Plotting the values (Figure 5)

suggest no clear differences in the variances between the two populations because there is a similar spread of latencies.

Table 4. Contrasts (indicated by "diff") between populations for the exploration measure: latency to approach within 20 cm of a novel environment. Note that “phi” in this table refers to a term in the gamma poisson model and not to what we refer to as the phi parameter in reversal learning.

	Mean	Standard deviation	Lower 89 percentile compatibility interval (5.5%)	Upper 89 percentile compatibility interval (94.5%)
Woodland	1697.40	229.76	1368.91	2058.43
Tempe	1137.56	181.84	875.64	1448.64
phi	1.59	0.29	1.15	2.09
diff_12	559.84	285.99	103.84	1017.56

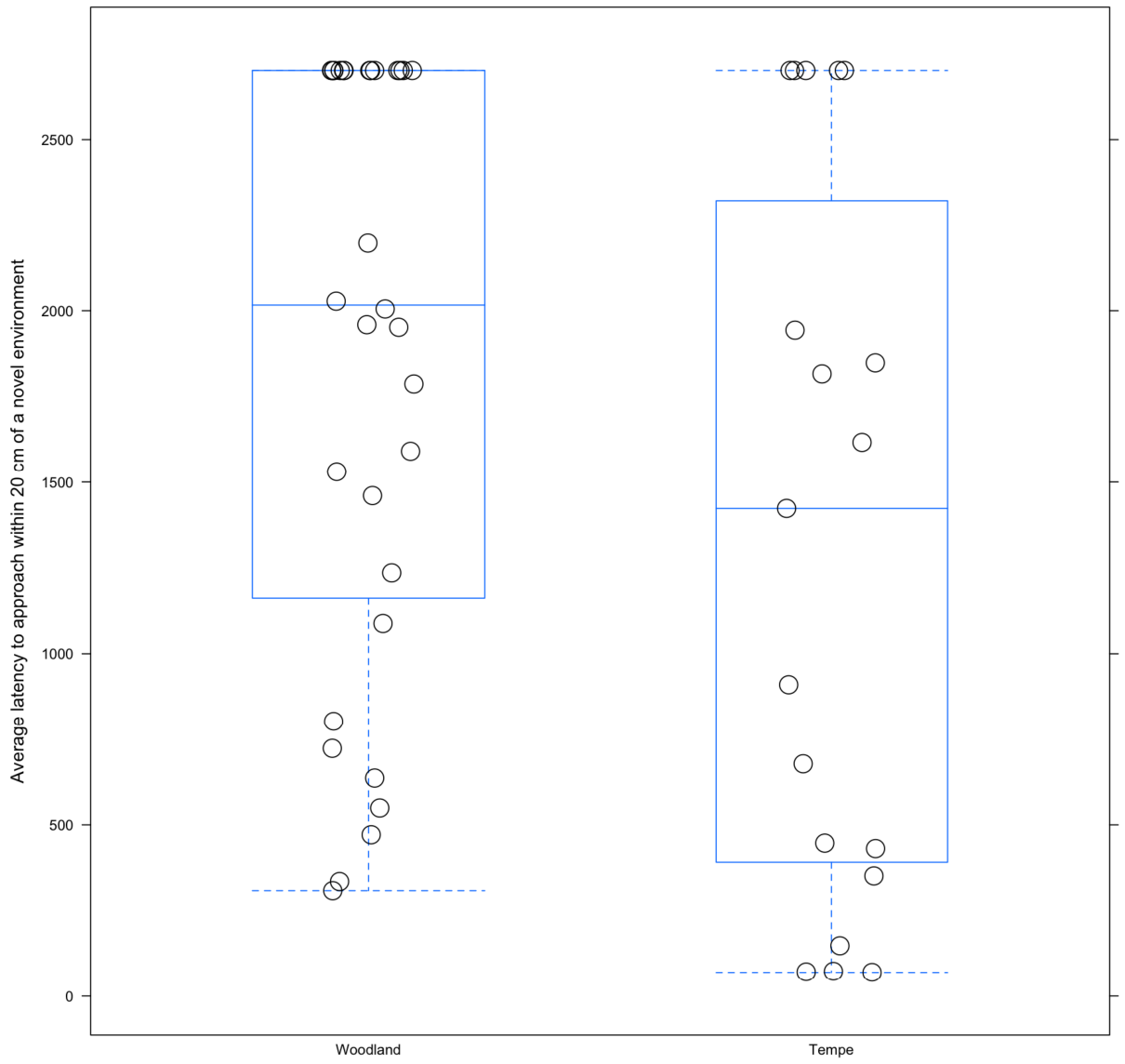


Figure 5. Average latency to approach within 20 cm of a novel environment in the exploration assay per individual at each site (n=32 Woodland, n=19 Tempe). Note that if an individual does not approach within 20 cm of the novel environment at Time 1 or 2, they are given a ceiling value of 2701, which is one second longer than the session length. The boxplots show the minimum, maximum, lower and upper quartiles, and median values. The black circles are the raw data from each individual.

Persistence

There are no strong site differences for persistence quantified as the proportion of trials participated in across the reversal and multiaccess box experiments (Figure 6). We would need a difference of more than 0.08 in the proportion of trials participated in to detect a difference between the sites (based on our power analysis in Supplementary Material 2, summarized in

Table SM1). However, the proportion differs by only 0.08 (Woodland=0.72, Tempe=0.80), and the site differences are unlikely to be larger than 0.08 (Table 5). Visual interpretation, through plotting the values (Figure 6), could suggest that the variance in persistence might be larger among the individuals in Woodland compared to Tempe because some of the Woodland individuals show lower persistence values than those in the Tempe individuals. We conducted an UNREGISTERED ANALYSIS which finds no support that the variances differ between the two populations (Levene's test for homogeneity of variance: df=1, F value=1.9, p=0.17).

Table 5. Contrasts (indicated by “diff”) between populations for the persistence measure: proportion of trials participated in across the reversal and multiaccess box experiments.

	Mean	Standard deviation	Lower 89 percentile compatibility interval (5.5%)	Upper 89 percentile compatibility interval (94.5%)
Woodland	0.78	0.01	0.77	0.80
Tempe	0.79	0.01	0.78	0.80
diff_12	0.00	0.01	-0.02	0.01

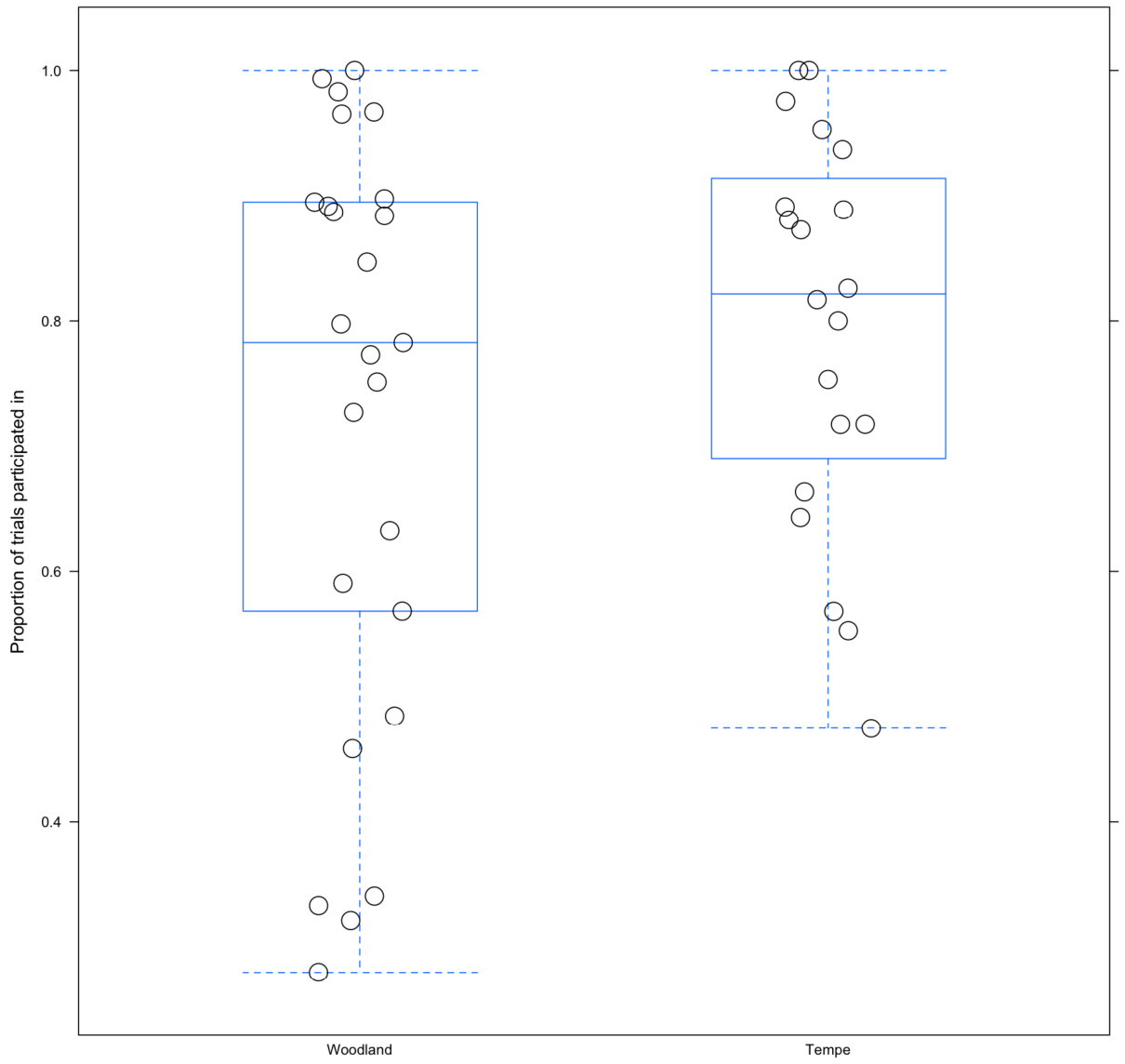


Figure 6. The proportion of trials participated in across the reversal and multiaccess box experiments is the measure of persistence per individual at each site (n=25 Woodland, n=20 Tempe). The boxplots show the minimum, maximum, lower and upper quartiles, and median values. The black circles are the raw data from each individual.

Repeatability of exploration and persistence

Exploration of the novel environment is repeatable in the Woodland population (current study: likelihood ratio test: R=0.70, p=0.001, confidence interval=0.2-1.0). Our previous analysis found that novel environment exploration was repeatable in the Tempe [@mccune2019exploration: R=0.72, p<0.001, confidence interval=0.42-0.88] grackles. Persistence is repeatable across both populations (likelihood ratio test: R=0.24, p=0.03, confidence interval=0.03-0.46).

DISCUSSION

We conduct behavioral experiments with great-tailed grackles from two populations: an older population in the middle of the expansion front (Tempe, Arizona), and a more recent population on the northern edge of their expansion in Woodland, California. Our measures of flexibility [using serial reversals in the Tempe population; @mccune2022flexmanip], exploration [Tempe: @mccune2019exploration; Woodland: reported here], and persistence (both populations reported here) are repeatable and show large inter-individual variation, which validates that these are stable traits that can be meaningfully compared. We find that individuals in the edge population are more innovative and less exploratory than the population in the middle of the expansion front, and that there are no population differences in behavioral flexibility or persistence. This supports the hypothesis that changes in particular behavioral traits are potentially important for facilitating a species' rapid geographic range expansion.

We find no support for the hypothesis that flexibility plays an important role in rapid geographic range expansions [e.g., @lefebvre1997feeding; @griffin2014innovation; @chow2016practice; @sol2000behavioural; @sol2002behavioural; @sol2005big; @sol2007big; @wright2010behavioral]. The finding that flexibility is not higher among individuals at the edge of the expansion range indicates that flexibility is not a latent trait that is called upon when individuals move into new areas.

It is possible that behavioral flexibility facilitated the increase of this species' habitat breadth beyond marshes when humans started to modify the environment thousands of years ago [@christensen2000fifteenth]. Great-tailed grackles are now almost exclusively associated with human modified environments [@selander1961analysis, @johnson2001great, @wehtje2003range], and when planning study sites, we initially wanted to compare forest versus urban grackle populations. However, we are unable to find a population that exclusively exists in forests (based on eBird.org data, Logan, pers. obs.). In another article produced from the same preregistration, @logan2020xpop, as the current article, we investigate the role of increased habitat availability in geographic range expansions by comparing rapidly expanding great-tailed grackles with their closest relative that is not rapidly expanding its range, boat-tailed grackles (*Q. major*) [@summers2023xpop]. We predict that great-tailed grackles expanded their range because suitable habitat (i.e., human modified environments) increased (prediction 1 alternative 1 in the preregistration). Results show that, between 1979 and 2019, great-tailed grackles increased their habitat breadth to include more urban, arid environments. In contrast, boat-tailed grackles moved into new suitable habitat that was made available by climate change. These results support the possibility that flexibility played a role in the ability to increase habitat breadth. We are currently conducting a behavioral flexibility experiment in boat-tailed grackles to determine whether they are less flexible than great-tailed grackles, which would further support the hypothesis that flexibility was involved in the great-tailed grackle rapid range expansion [in the same preregistration as the current study: @logan2020xpop]. Unfortunately, we discovered in our first boat-tailed grackle field season in 2022 that they do not do well in captivity. Consequently, we will not continue the aviary tests in this species. Therefore, we only have comparable data from the aviary tests for two (reversal), four (multiaccess box), and five

(persistence) individuals. The boat-tailed grackle exploration videos are not coded and therefore not included in the analysis. Although the boat-tailed grackle sample size is too small to arrive at robust conclusions, we analyze their data here to give an indication of useful directions for future research. We find that boat-tailed grackles have **similar levels of flexibility** as both populations of great-tailed grackles; boat-tailed grackles are **less innovative** than the Woodland, but not the Tempe great-tailed grackles; and boat-tailed grackles are **less persistent** than both great-tailed grackle populations (see model outputs in Supplementary Material 4). This suggests that we might not find differences in flexibility between the two species. However, we are currently conducting reversal learning experiments in the wild in both species to determine whether this is a robust result [following the hypotheses in the preregistration at @logan2020xpop and methods and analyses in @logan2022manyindividuals].

The ability of great-tailed grackles to move into new habitats might be a species specific ability that has been ongoing for many years, which could be linked to the high levels of flexibility in this species being relatively fixed [@wright2010behavioral]. great-tailed grackles are flexible on the reversal learning task and are perhaps at their upper limit uniformly across their range. With an average reversal learning speed of 74 trials (using the data in the current article), great-tailed grackles are as flexible as great (*Parus major*) and blue (*Cyanistes caeruleus*) tits [average 59 trials; @morand2022cognitive] and three species of Darwin's finches (average 89 trials); and more flexible than Pinyon jays (average 155 trials), Clark's nutcrackers (average 143 trials), California scrub jays (average 191 trials), pigeons (average 168 trials) [data reported in @tebbich_tale_2010, but not in the original articles: @bond2007serial and @lissek2002impaired], and mice [average approximately 150 trials, @laughlin2011genetic]. Perhaps great-tailed grackles maintain a high level of flexibility across their range in response to daily changes in their local environment (e.g., the changing schedules of cafes with outdoor seating areas and garbage pick up times), rather than specifically in response to larger changes that might occur less frequently (e.g., traveling farther to exploit new foraging opportunities or moving to a new area).

Another alternative is that we measured the edge population too long after their initial establishment, during which time they potentially exhibited more flexibility for their initial adaptation phase to the new area [@wright2010behavioral]. If the sampled individuals had already been living at this location for long enough (or for their whole lives) to have learned what they need to about this particular environment (e.g., there may no longer be evidence of increased flexibility/innovativeness/exploration/persistence), there may be no reason to maintain population diversity in these traits to continue to learn about this environment. In this case, because differences in innovativeness are found, this trait could have different timing in the process of establishing in a new location (i.e., be required for longer). Great-tailed grackles occur more irregularly in areas further north of our edge site, and flexibility might be higher in more northern individuals from areas where stable populations are not yet established. However, evidence from experimental evolution suggests that, even after 30 generations there is no change in exploration of a novel environment or other behaviors (aggression, social grooming, courtship, and orientation) when comparing domestic guinea pigs with 30 generations

of wild-caught captive guinea pigs [@kunzl2003wild], whereas artificial selection can induce changes in spatial ability in as little as two generations [@kotrschal2013artificial]. This means it is likely that we would have detected population differences if such differences were linked with adapting to a new environment.

Differences in innovativeness and exploration are associated with the great-tailed grackle's rapid geographic range expansion. An increase in innovation in newly established populations can facilitate innovating new foraging techniques and exploiting new food sources, while a decrease in exploration can reduce their risk of encountering danger in a new area. The relatively little evidence from invasive species that are also expanding their geographic ranges shows similar results. Common mynas (*Acridotheres tristis*) on the invasion front are more innovative than those from populations away from the front and in their native range [@cohen2020innovation], and spiders from edge populations are less exploratory than those from core populations [@chuang2021personality]. While great-tailed grackles are not considered an invasive species because they introduced themselves rather than being introduced by humans, comparing them with invasive species is useful because the dynamics after the introduction stage should be similar (i.e., establishing in a new area and spreading out from there) [@chapple2012can]. Note that wild great-tailed grackles were caught from north of Rio de la Antigua, Mexico by the Aztec emperor, Auitzotl (1486-1502), and introduced approximately 370 km inland to the Valley of Mexico (Tenochititlan & Tlatelolco) where they reproduced and spread [@haemig2011introduction; @haemig2012introduction; @haemig2014aztec]. By 1577, they spread at least 100 km including back to their native range [@haemig2011introduction]. This indicates that great-tailed grackles had already spread this far north by themselves before the introduction at a parallel latitude, and that they continued their spread without the help of human-facilitated introductions.

Flexibility is causally related with innovativeness in great-tailed grackles [@logan2023flexmanip, measured on the Tempe individuals included in the current study]. We manipulated flexibility in the Tempe grackles by giving a manipulated group serial reversals until they passed quickly. The manipulated grackles were then given an innovation test (the multiaccess box) and found to be more innovative (solved more loci) compared to control grackles who only experienced one reversal. Flexibility, the ability to recognize that something about the environment has changed and decide to consider other options for deploying behavior [@mikhalevich_is_2017], is distinct from innovation, which is the specific stringing together of particular behaviors in response to the decision to change behavior in some way [@griffin2014innovation]. That they are causally related does not mean that they must always be associated to the same degree because there can be other variables that additionally influence one or both traits differentially across time and space (e.g., environmental unpredictability, features of the items they forage on that differ and require different access methods). We are currently investigating how flexibility and exploration, and flexibility and a different measure of persistence (number of functional and/or non-functional touches to test apparatuses) are related in the Tempe grackles [@mccune2019exploration]. Additionally, we are determining to what extent the aviary measure of exploration is a proxy for how the Tempe individuals use space in the wild after their release [@mccune2020spaceuse].

In conclusion, rather than flexibility being associated with a rapid geographic range expansion, as is widely hypothesized, we find that higher innovation and lower exploration levels are the key behavioral traits associated with the great-tailed grackle's edge population in comparison with an older population closer to the original range. The term "behavioral flexibility" is defined and measured in a variety of ways in the literature (or it is not defined at all). For example, the detour task (individuals must walk around a transparent barrier to access a food reward) is sometimes considered a test of flexibility, sometimes a test of self control, and sometimes a test of both. However, theoretically and empirically it measures a trait that is not, and is not related to, flexibility or self control, but rather a different trait: motor inhibition [@logan2021inhibition]. We argue that calling many types of traits "flexibility" without proper (or sometimes any) theoretical justification and without validating methods is detrimental because it confounds our ability to answer questions about the broader significance of flexibility and how it is genuinely involved in large scale changes [@logan_beyond_2017; @mikhalevich_is_2017]. Our research program shows the value of clearly defining terms for behavioral traits, validating the methods intended to measure those traits, and understanding how certain traits relate to each other (causally if possible) before attempting to answer broader cross population questions.

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], SP606267 [2018], SP639866 [2019], and SP402153 [2020])
4. Institutional Animal Care and Use Committee at Arizona State University (protocol number 17-1594R)
5. California Department of Fish and Wildlife (scientific collecting permit [specific use] number S-192100001-19210-001)
6. RegionalSan (access permit number AP 2021-01)

AUTHOR CONTRIBUTIONS

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

McCune: Method development, data collection, data analysis and interpretation, revising/editing.

LeGrande-Rolls: Data collection, revising/editing.

Marfori: Data collection, revising/editing.

Hubbard: Data collection, revising/editing.

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

FUNDING

This research is funded by the Department of Human Behavior, Ecology and Culture at the Max Planck Institute for Evolutionary Anthropology.

CONFLICT OF INTEREST DISCLOSURE

We, the authors, declare that we have no financial conflicts of interest with the content of this article. CJ Logan and D Lukas are Recommenders at PCI Ecology, and CJ Logan used to be on the Managing Board at PCI Ecology (2018-2022).

ACKNOWLEDGEMENTS

We thank Kristine Johnson for technical advice on great-tailed grackles; Julia Cissewski and Sophie Kaube 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; Alexis Breen and Vincent Kiepsch for video coding; Ann Brice, Woodland-Davis Clean Water Agency, Regional San, and Conaway Ranch for hosting the research on their land; and Rhonda Oates and the vet team at UC Davis for veterinary consultations.

SUPPLEMENTARY MATERIAL 1: Sample size rationale

We summarize the minimum sample sizes and their associated detection limits in Table SM1, which allowsed us to determine whether populations are different from each other (detailed in the Analysis section for each experiment).

Table SM1. A summary of the measure of interest in each experiment, the distribution used for the analysis, the minimum detectable difference between site means, and the minimum sample size that goes with the minimum detectable difference.

Code

Experiment	Measurement	Distribution	Minimum difference between site means	Minimum sample size
Reversal	Phi (learning rate)	Gamma	Differences of 0.01 are likely to be detected (based on models with 20 individuals per site, however this is likely to hold for the the minimum sample size as well) (Figures SM2.15 and SM2.26)	15

Reversal	Lambda (random choice rate)	Gamma	Differences of 3 are likely to be detected (based on models with 20 individuals per site, however this is likely to hold for the the minimum sample size as well) (Figures SM2.1~~5~~ and 2.2~~6~~)	15
Multiaccess box	Number of loci solved	Binomial	Differences of 1.2 loci are likely to be detected (Table SM3~~3~~)	15
Exploration	Latency to approach novel object	Gamma-Poisson	Differences of at least 450 seconds are likely to be detected (Table SM2~~3~~)	15
Persistence	Percent of trials participated in	Normal	Difference of at least 0.08 in the proportion of trials participated in (Table SM2~~3~~)	18

SUPPLEMENTARY MATERIAL 2: Simulations for power analyses

Hypothesis-specific mathematical model

Following procedures in McElreath (2016), we construct~~ed~~ a **hypothesis-appropriate mathematical model** for each of the response variables that examines differences in the response variable between sites (each site represents a grackle population). These models take the form of:

$$y \sim \alpha[site]$$

y is the response variable (flexibility, innovation, exploration, or persistence). There is~~will be~~ one intercept, α, per site and we ~~will~~ estimate the site's average and standard deviation of the response variable.

We formulate~~d~~ these models in a Bayesian framework. We determine~~d~~ the priors for each model by performing prior predictive simulations based on ranges of values from the literature to check that the models are covering the likely range of results.

We ~~will~~ then perform pairwise contrasts to determine at what point we can~~will~~ be able to detect differences between sites by manipulating sample size, and α means, and standard deviations. Before running the simulations, we decided that a model would detect an effect if 89% of the difference between two sites is on the same side of zero (following McElreath (2016)). We are using a Bayesian approach, therefore comparisons are based on samples from the posterior distribution. We ~~will~~ draw 10,000 samples from the posterior distribution, where each sample ~~will~~ has~~ve~~ an estimated mean for each population. For the first contrast, within each sample, we subtract the estimated mean of the edge population from the estimated mean of the core population. For the second contrast, we subtract the estimated mean of the edge population from the estimated mean of the middle population. For the third contrast, we subtract the estimated mean of the middle population from the estimated mean of the core population. We ~~will~~ now have samples of differences between all of the pairs of sites, which we can use to assess whether any site is systematically larger or smaller than the others. We ~~will~~ determine whether this is the case by estimating what percentage of each sample of differences is either larger or smaller than zero. For the first contrast, if 89% of the differences are larger than zero, then the core population has a larger mean. If 89% of the differences are smaller than zero, then the edge population has a larger mean.

Flexibility analysis

Power analyses: We also use~~d~~ the simulations to estimate our ability to detect differences in ϕϕ and λλ between sites based on extracting samples from the posterior distribution. We ru~~a~~n two different sets of simulations: ~~the first set of simulations recreated choices for 20 birds per population across initial learning and reversal trials; the second set of simulations~~ we first sample~~d~~ between 9 and 24 birds from populations with pre-specified ϕϕ and λλ means to determine the minimum sample size required to detect whether two populations are different. This set of simulations shows how different site sample sizes change detection levels: once a sample size of 15 is reached, there are only minimal differences in detection abilities compared to larger sample sizes (Figure SM2.1~~446~~). The second set of simulations recreates~~recreatsed~~ choices for 20 birds per population across initial learning and reversal trials from which we estimate~~d~~ their ϕϕ and λλ. We ~~chose to~~ simulate 20 birds per population because this number is above the threshold we detected in the first set of simulations and it appears~~ed~~ a feasible sample size. We expect~~ed~~ that the noise in the probabilistic choices of individuals might reduce the differences that can be detected compared to the first simulation where ϕϕ and λλ a~~we~~re assumed to be exactly known for each individual. This second ~~e first~~ set of simulations shows~~ed~~ that we have a very high chance of detecting that two sites are different from each other if the difference in their ϕϕ is 0.01 or greater and/or if the difference in their λλ is 3 or greater, based on data from 20 simulated individuals per site (Figure SM2.2~~35~~). ~~The second set of simulations shows how different site sample sizes change detection levels: once a~~

~~sample size of 15 is reached, there are only minimal differences in detection abilities compared to larger sample sizes (Figure 46).~~ It appears that there is more variability in the $\lambda\lambda$ estimates for each bird based on their choices, meaning that with the learning model, which estimates $\lambda\lambda$ from the choices, the differences between sites have to be larger (than if we were able to infer lambda directly) to be reliably detected. Given that we have to infer $\phi\phi$ and $\lambda\lambda$ from the choices~~The second set of simulations assumed that we could infer $\lambda\lambda$ directly, however it is not possible to directly measure $\lambda\lambda$. Therefore~~, the power curves in Figure SM2~~3~~.1~~5~~ are more reliable than those in Figure SM2.2~~46~~.

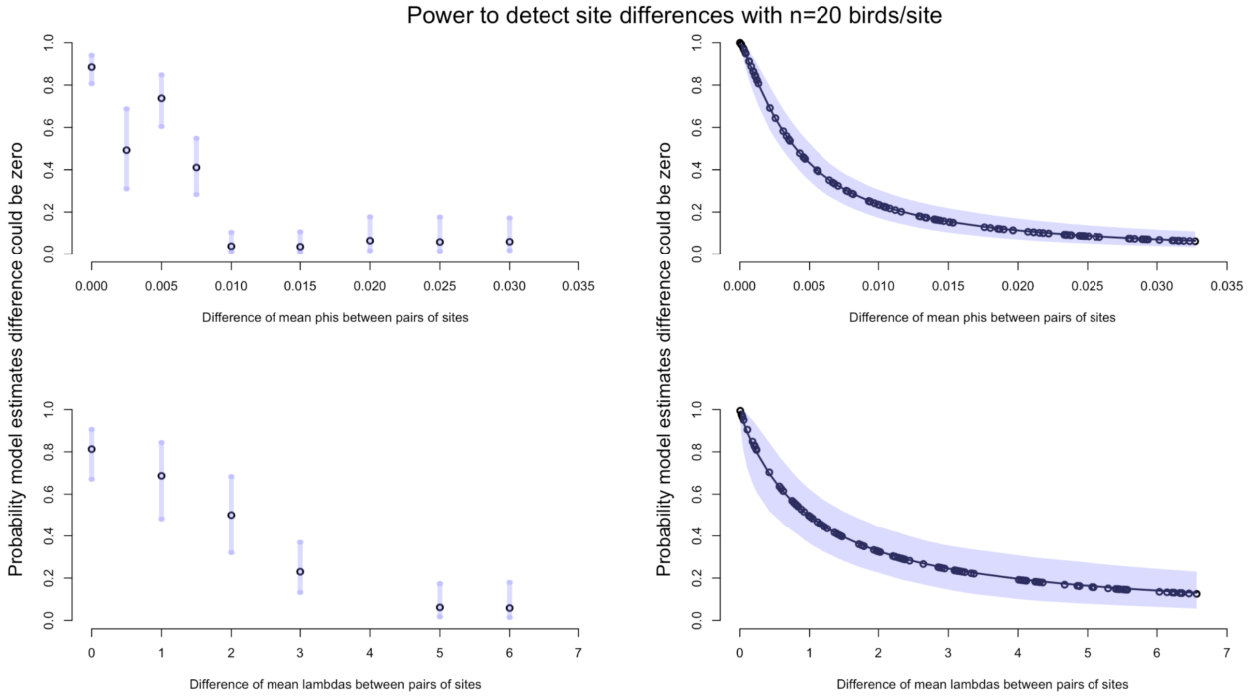
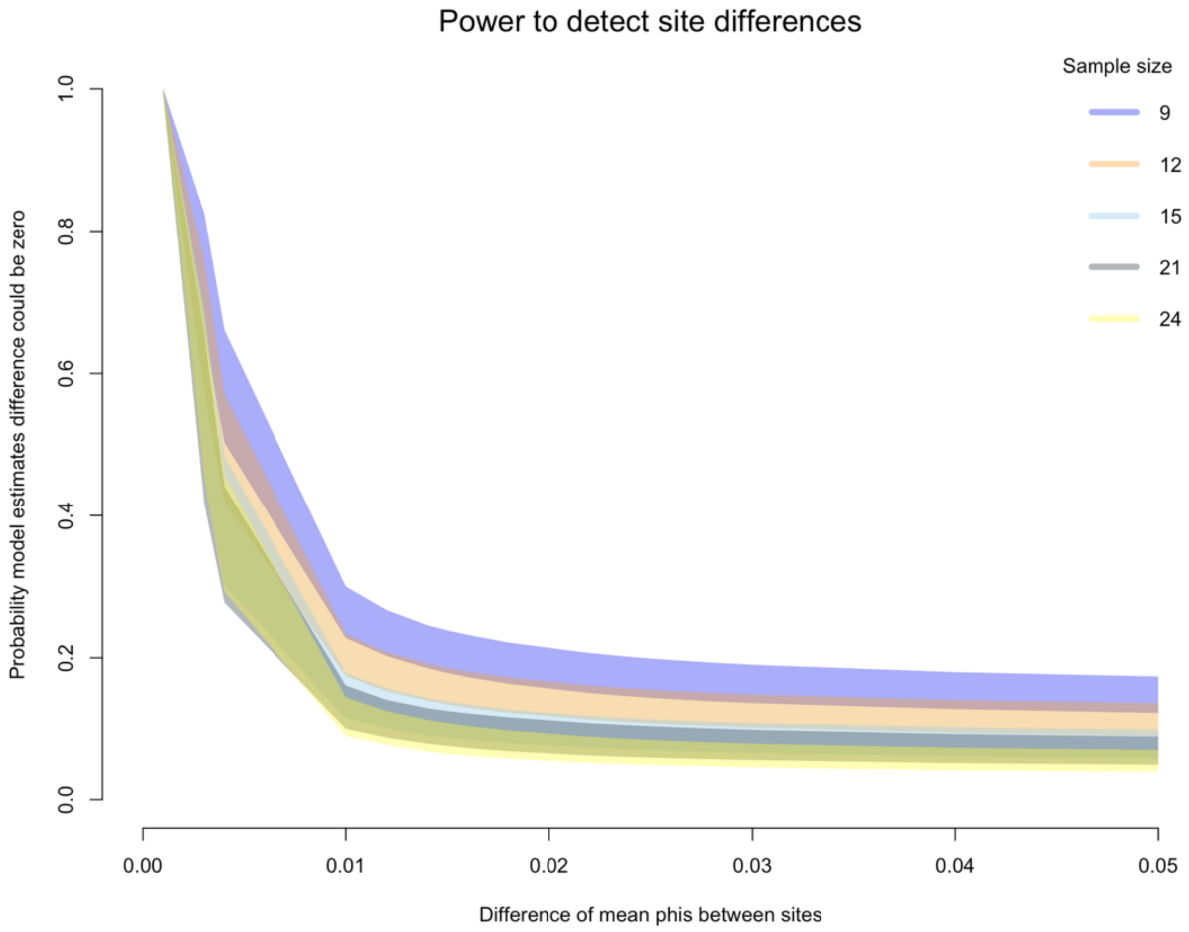


Figure SM2.1~~35~~. How small of a site difference in phi and lambda can we detect? The probability that the model estimates that the difference shown on the x axis is zero, meaning that the model assumes that it is possible that these two estimates come from a population with the same phi or lambda. Each point is the mean phi or mean lambda from one site minus the mean phi or mean lambda from another site (calculated from 20 individuals per site) for all pairwise comparisons for all 32 simulated sites (for a total of 496 pairwise comparisons). Left panels: error bars=89% compati~~a~~bility intervals. Right panels: shaded areas=97% prediction intervals.


Power to detect site differences
Sample size
9
12
15
21
24
Probability model estimates difference could be zero
0.0
0.2
0.4
0.6
0.8
1.0
0.00
0.01
0.02
0.03
0.04
0.05
Difference of mean phis between sites

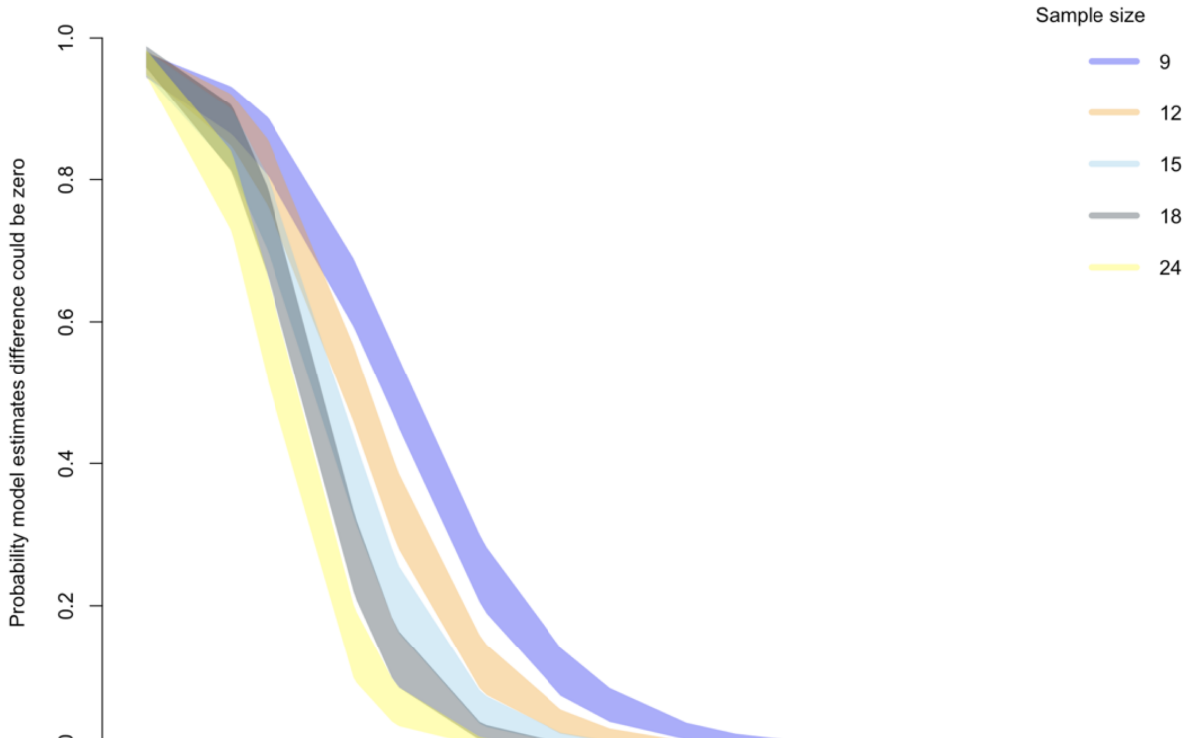

Sample size
9
12
15
18
24
Probability model estimates difference could be zero
0.2
0.4
0.6
0.8
1.0

Figure SM2.246. How do detection differences vary according to sample size differences? The probability that the model estimates that the difference shown on the x axis is zero, meaning that the model assumes that it is possible that these two estimates come from a population with the same phi or lambda. The x-axis is the mean phi or mean lambda from one site minus the mean phi or mean lambda from another site for all pairwise comparisons for all 14 sites (for a total of 91 pairwise comparisons). Each shaded region is the 97% prediction interval for that particular sample size.

Innovation analysis

After building the model (see Methods), wWe then run the **mathematical model** and performed pairwise contrasts and determined that we are be able to detect differences between sites with a sample size of 15 at each site if the average number of loci solved differs by 1.2 loci or more and the standard deviation is generally a maximum of 0.9 at each site (Table SM3). For a sample size of 20 at each site, we are be able to detect site differences if the average number of loci solved differs by 0.7 of a locus or more and the standard deviation is generally a maximum of 1 at each site (Table SM3). Note: the Arizona sample size is 11 for the multiaccess log and 17 on a similar multiaccess box.

Because the mean and the variance are linked in the binomial distribution, and because the variance simulations in the flexibility analysis showed that we arewill not be able to robustly detect differences in variance between sites, we will plot the variance in the number of loci solved between sites to determine whether the edge population has a wider or narrower spread than the other two populations.

Exploration analysis

After building the model (see Methods), wWe then ruan the **mathematical model** and performed pairwise contrasts and determined that we arewill be able to detect differences between sites with a potential sample size of 15 at each site or 20 at each site if the average latency to approach the novel environmentobject differs by at least 450 sec between at each sites (Table SM3). We keept the shape of the curve (which can be thought of as similar to a standard deviation or the variance) the same across sites because we do not think this assumption will change across populations (i.e., there will be lots of variation at each site with some individuals approaching almost immediately, others in the middle of the session, and others near the end).

Because the mean and the variance are linked in the gamma-Poisson distribution, and because the variance simulations in the flexibility analysis showed that we will not be able to robustly detect differences in variance between sites, we will plot the variance in the latency to approach the novel environmentobject between sites to determine whether the edge population has a wider or narrower spread than the other two populations.

Persistence analysis

After building the model (see Methods), wWe then ruan the **mathematical model** and performed pairwise contrasts and determined that we arewill be able to detect differences between sites with a potential sample size of 15 per site or 20 per site if the average proportion of trials participated in differs by at least 0.08 and the standard deviation is generally a maximum of 0.25 at each site (Table SM33).

Because the mean and the variance are linked in the binomial distribution, and because the variance simulations in the flexibility analysis showed that we arewill not be able to robustly detect differences in variance between sites, we will plot the variance in the proportion of trials participated in between sites to determine whether the edge population has a wider or narrower spread than the other two populations.

Table SM23. Simulation outputs from **varying sample size (n), and αα means and standard deviations**. We calculate pairwise contrasts between the estimated means from the posterior distribution: if for a large sample the difference is both positive and negative and crosses zero (yes),

then we are not able to detect differences between the two sites. If the differences between the means are all on one side of zero for 89% of the posterior samples (no), then we are able to detect differences between the two sites. We chose the 89% interval based on (McElreath, 2016). Note that for latency, there is no mu_sd, but rather one phi that is the same for all sites. mu=average, sd=standard deviation. The numbers 1-3 in the column titles refer to sites 1-3 as do S1-3 (the simulations were run on a total of three sites because we originally planned to collect data at two to three sites), mu=average, sd=standard deviation. Loci solved is the innovativeness measure, latency is the exploration measure, and trials participated in is the persistence measure.

Response variable	n	mu site 1	mu site 2	mu site 3	mu_sd site 1	mu_sd site 2	mu_sd site 3	Difference crosses zero? Sites 1 vs 2	Difference crosses zero? Sites 1 vs 3	Difference crosses zero? Sites 2 vs 3	Notes
loci solved	60	1.90	2.10	3.60	0.50	0.50	0.50	No	No	No	
loci solved	60	1.90	2.10	2.20	0.50	0.50	0.50	Yes	Yes	Yes	
loci solved	60	1.90	2.10	2.30	0.50	0.50	0.50	Yes	Yes	Yes	
loci solved	60	1.90	2.10	3.50	0.50	0.50	0.50	No	No	No	
loci solved	60	1.90	2.10	3.00	0.50	0.50	0.50	Yes	No	No	
loci solved	60	1.90	2.10	2.80	0.50	0.50	0.50	Yes	No	Yes	
loci solved	60	1.80	2.10	3.00	0.50	0.50	0.50	Yes	Yes	No	
loci solved	60	2.00	2.50	3.00	0.50	0.50	0.50	No	Yes	No	
loci solved	60	2.00	2.50	3.10	0.50	0.50	0.50	No	No	Yes	
loci solved	60	1.90	2.50	3.20	0.50	0.50	0.50	Yes	No	No	
loci solved	60	1.80	2.50	3.30	0.50	0.50	0.50	No	No	Yes	
loci solved	60	1.70	2.50	3.40	0.50	0.50	0.50	No	No	No	
loci solved	60	1.70	2.50	3.40	1.00	1.00	1.00	No	No	No	
loci solved	60	1.70	2.50	3.40	1.50	1.50	1.50	Yes	No	No	
loci solved	60	1.70	2.50	3.40	1.30	1.30	1.30	Yes	Yes	Yes	
loci solved	60	1.00	2.00	3.00	0.50	0.50	0.50	No	No	Yes	
loci solved	60	1.00	2.00	3.00	0.50	0.50	0.50	No	No	No	
loci solved	60	1.00	2.00	3.00	0.30	0.40	0.50	No	No	No	
loci solved	60	1.00	2.00	3.00	0.60	0.70	0.50	No	No	No	
loci solved	60	1.00	2.00	3.00	0.70	0.70	0.70	No	No	No	
loci solved	60	1.00	2.00	3.00	0.90	0.90	0.90	No	No	No	
loci solved	60	1.00	2.00	3.00	1.00	1.00	1.00	No	No	No	

SUPPLEMENTARY MATERIAL 3: Interobserver reliability of dependent variables

To determine whether experimenters coded the dependent variables in a repeatable way, hypothesis-blind video coders were first trained in video coding the dependent variables (reversal learning and multiaccess log: whether the bird made the correct choice or not; exploration: latency to approach), requiring a Cohen's unweighted kappa (reversal and multiaccess categorical variables) or an intra-class correlation coefficient (ICC; exploration continuous variable) of 0.90 or above to pass training. This threshold indicated that the two coders (the experimenter and the video coder) agreed with each other to a high degree (kappa: Landis & Koch (1977); ICC: Hutcheon et al. (2010)). After passing training, the video coders coded 20% of the videos for each experiment (except for exploration for which 15% of the videos were coded due to an unexpectedly high sample size for this assay). T~~and t~~he kappa and ICC were calculated to determine how objective and repeatable scoring was for each variable, while noting that the experimenter has the advantage over the video coder because watching the videos is not as clear as watching the bird participate in the trial from the aisle of the aviaries. The unweighted kappa was used when analyzing a categorical variable where the distances between the numbers are meaningless (0=incorrect choice, 1=correct choice, -1=did not participate), and the ICC was used for continuous variables where distances are meaningful (e.g., if coders disagree by a difference of 2 s rather than 5 s, this is important to account for).

Interobserver reliability training

To pass **interobserver reliability (IOR) training**, video coders needed an ICC or Cohen's unweighted kappa score of 0.90 or greater to ensure the instructions were clear and that there was a high degree of agreement across coders. Video coders, Alexis Breen and Vincent Kiepsch, passed interobserver reliability training for exploration in a previous article (McCune ~~KB~~ et al., 2019) where their training results can be found.

Lea Gihlein (compared with experimenter's live coding):

- Reversal learning: correct choice unweighted Cohen's Kappa=1.00 (confidence boundaries=1.00-1.00, n=21 data points)
- Multiaccess box: correct choice unweighted Cohen's Kappa=1.00 (confidence boundaries=1.00-1.00, n=29 data points)
- Multiaccess box: correct choice unweighted Cohen's Kappa=1.00 (confidence boundaries=1.00-1.00, n=29 data points)

Interobserver reliability

Interobserver reliability scores (minimum 20% of the videos) were as follows:

Lea Gihlein (compared with experimenter's live coding):

- Reversal learning (5/19 birds): correct choice unweighted Cohen's Kappa=1.00 (confidence boundaries=0.99-1.00, n=707 data points)
- Multiaccess box (5/23 birds): correct choice unweighted Cohen's Kappa=0.92 (confidence boundaries=0.81-1.00, n=63 data points)
- Multiaccess box (5/23 birds): locus solved unweighted Cohen's Kappa=1.00 (confidence boundaries=1.00-1.00, n=48 data points)

Vincent Kiepsch (compared with Breen's video coding):

- Exploration (5/34 birds): latency to land on the ground unweighted Cohen's Kappa=0.998 (confidence boundaries=0.997-0.999, n=32 data points)

SUPPLEMENTARY MATERIAL 4: boat-tailed grackle model outputs

Table SM4. Results for the comparison between the boat-tailed grackle (BTGR) population in Lake Placid and Venus, Florida and the great-tailed grackle populations in Tempe, Arizona and Woodland, California. Contrasts (indicated by "diff") between populations show whether there was a difference (compatibility interval does not cross zero) or not (compatibility interval crosses zero) for that pair of populations. Populations are labeled as follows: 1=boat-tailed grackles (BTGR), 2=Woodland great-tailed grackles, 3=Tempe great-tailed grackles (e.g., diff_12 means that BTGR and Woodland are being compared).

	Mean	Standard deviation	Lower 89 percentile compatibility interval (5.5%)	Upper 89 percentile compatibility interval (94.5%)
FLEXIBILITY	NA	NA	NA	NA
BTGR phi	0.03	0.01	0.02	0.04
BTGR lambda	4.51	1.34	3.11	5.93
diff_12 phi	0.00	0.01	-0.01	0.01
diff_12 lambda	0.23	0.97	-1.06	1.97
diff_13 phi	0.00	0.01	-0.01	0.01
diff_13 lambda	0.43	1.01	-0.79	2.32
	NA	NA	NA	NA
INNOVATIVENESS	NA	NA	NA	NA
BTGR	0.36	0.11	0.19	0.53
Woodland	0.76	0.04	0.69	0.83
Tempe	0.50	0.06	0.40	0.60
diff_12	-0.41	0.12	-0.59	-0.22
diff_13	-0.14	0.13	-0.34	0.07
	NA	NA	NA	NA
PERSISTENCE	NA	NA	NA	NA
BTGR	0.69	0.02	0.66	0.72
Woodland	0.78	0.01	0.77	0.79
Tempe	0.79	0.01	0.78	0.80
diff_12	-0.10	0.02	-0.13	-0.06
diff_13	-0.10	0.02	-0.13	-0.06

REFERENCES

Auersperg, A. M. I., Bayern, A. M. P. von, Gajdon, G. K., Huber, L., & Kacelnik, A. (2011). Flexibility in problem solving and tool use of kea and New Caledonian crows in a multi access box paradigm. *PLOS ONE*, *6*(6), e20231. https://doi.org/10.1371/journal.pone.0020231

Auersperg, A. M. I., Szabo, B., Von Bayern, A. M., & Kacelnik, A. (2012). Spontaneous innovation in tool manufacture and use in a goffin's cockatoo. *Current Biology*, *22*(21), R903–R904.

Bird, C. D., & Emery, N. J. (2009). Insightful problem solving and creative tool modification by captive nontool-using rooks. *Proceedings of the National Academy of Sciences*, *106*(25), 10370–10375.

Blaisdell, A., Seitz, B., Rowney, C., Folsom, M., MacPherson, M., Deffner, D., & Logan, C. J. (2021). *Do the more flexible individuals rely more on causal cognition? Observation versus intervention in causal inference in great-tailed grackles (version 5 of this preprint has been peer reviewed and recommended by peer community in ecology [https://doi.org/10.24072/pci.ecology.100076])*. https://doi.org/10.31234/osf.io/z4p6s

Bolker, B. M. (2008). *Ecological models and data in r*. Princeton University Press.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*(1).

Chejanovski, Z. A., Avilés-Rodríguez, K. J., Lapiedra, O., Preisser, E. L., & Kolbe, J. J. (2017). An experimental evaluation of foraging decisions in urban and natural forest populations of anolis lizards. *Urban Ecosystems*, *20*(5), 1011–1018.

Chow, P. K. Y., Lea, S. E., & Leaver, L. A. (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

Ciani, A. C. (1986). Intertroop agonistic behavior of a feral rhesus macaque troop ranging in town and forest areas in india. *Aggressive Behavior*, *12*(6), 433–439.

Collias, E. C., & Collias, N. E. (1964). The development of nest-building behavior in a weaverbird. *The Auk*, *81*(1), 42–52.

Eddelbuettel, D., & François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, *40*(8), 1–18. https://doi.org/10.18637/jss.v040.i08

Federspiel, I. G., Garland, A., Guez, D., Bugnyar, T., Healy, S. D., Güntürkün, O., & Griffin, A. S. (2017). Adjusting foraging strategies: A comparison of rural and urban common mynas (acridotheres tristis). *Animal Cognition*, *20*(1), 65–74.

Gamer, M., Lemon, J., Gamer, M. M., Robinson, A., & Kendall's, W. (2012). Package "irr." *Various Coefficients of Interrater Reliability and Agreement*.

Goldewijk, K. K. (2001). Estimating global land use change over the past 300 years: The HYDE database. *Global Biogeochemical Cycles*, *15*(2), 417–433.

Griffin, A. S., & Guez, D. (2014). Innovation and problem solving: A review of common mechanisms. *Behavioural Processes*, *109*, 121–134. https://doi.org/10.1016/j.beproc.2014.08.027

Hadfield, J. (2010). MCMCglmm: Markov chain monte carlo methods for generalised linear mixed models. *Tutorial for MCMCglmm Package in R*, *125*.

Hadfield, J. (2014). *MCMCglmm course notes*. http://cran.r-project.org/web/packages/MCMCglmm/vignettes/CourseNotes.pdf

Hanski, I., & Gilpin, M. (1991). Metapopulation dynamics: Brief history and conceptual domain. *Biological Journal of the Linnean Society*, *42*(1-2), 3–16.

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

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

Kotrschal, A., Rogell, B., Bundsen, A., Svensson, B., Zajitschek, S., Brännström, I., Immler, S., Maklakov, A. A., & Kolm, N. (2013). Artificial selection on relative brain size in the guppy reveals costs and benefits of evolving a larger brain. *Current Biology*, *23*(2), 168–171.

Künzl, C., Kaiser, S., Meier, E., & Sachser, N. (2003). Is a wild mammal kept and reared in captivity still a wild animal? *Hormones and Behavior*, *43*(1), 187–196.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159–174.
Laumer, I., Call, J., Bugnyar, T., & Auersperg, A. (2018). Spontaneous innovation of hook-bending and unbending in orangutans (pongo abelii). *Scientific Reports*, *8*(1), 1–13.
Lefebvre, L., Whittle, P., Lascaris, E., & Finkelstein, A. (1997). Feeding innovations and forebrain size in birds. *Animal Behaviour*, *53*(3), 549–560. https://doi.org/10.1006/anbe.1996.0330
Liu, X., Huang, Y., Xu, X., Li, X., Li, X., Ciais, P., Lin, P., Gong, K., Ziegler, A. D., Chen, A., et al. (2020). High-spatiotemporal-resolution mapping of global urban change from 1985 to 2015. *Nature Sustainability*, 1–7.
Logan, C. J. (2016a). Behavioral flexibility and problem solving in an invasive bird. *PeerJ*, *4*, e1975. https://doi.org/10.7717/peerj.1975
Logan, C. J. (2016b). Behavioral flexibility in an invasive bird is independent of other behaviors. *PeerJ*, *4*, e2215.
Logan, C. J., Avin, S., Boogert, N., Buskell, A., Cross, F. R., Currie, A., Jelbert, S., Lukas, D., Mares, R., Navarrete, A. F., et al. (2018). Beyond brain size: Uncovering the neural correlates of behavioral and cognitive specialization. *Comparative Cognition & Behavior Reviews*.
Logan, C. J., MacPherson, M., Rowney, C., Bergeron, L., Seitz, B., Blaisdell, A., Folsom, M., Johnson-Ulrich, Z., & McCune, K. B. (2019). Is behavioral flexibility manipulatable and, if so, does it improve flexibility and problem solving in a new context? *In Principle Acceptance by PCI Ecology of the Version on 26 Mar 2019*. http://corinalogan.com/Preregistrations/g_flexmanip.html
Logan, CJ, McCune, KB, & Lukas, D. (2023). Data: Implementing a rapid geographic range expansion - the role of behavior changes. *Knowledge Network for Biocomplexity*, *Data package*.
Manrique, H. M., & Call, J. (2011). Spontaneous use of tools as straws in great apes. *Animal Cognition*, *14*(2), 213–226.
McCune, K. B. (2018). *Cognition gone wild: A test of the social intelligence hypothesis in wild birds* [PhD thesis].
McCune, K. B., Jablonski, P., Lee, S., & Ha, R. R. (2019). Captive jays exhibit reduced problem-solving performance compared to wild conspecifics. *Royal Society Open Science*, *6*(1), 181311.
McCune, KB, MacPherson, M, Rowney, C, Bergeron, L, Folsom, M, & Logan, C. (2019). Is behavioral flexibility linked with exploration, but not boldness, persistence, or motor diversity? *In Principle Acceptance by PCI Ecology of the Version on 27 Mar 2019*. http://corinalogan.com/Preregistrations/g_exploration.html
McElreath, R. (2016). *Statistical rethinking: A bayesian course with examples in r and stan*. CRC Press. https://doi.org/10.1201/9781315372495
McElreath, R. (2020). *Rethinking: Statistical rethinking book package*.
Mery, F., & Kawecki, T. J. (2005). A cost of long-term memory in drosophila. *Science*, *308*(5725), 1148–1148.
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*(5), 482–492.
Mikhalevich, I., Powell, R., & Logan, C. (2017). Is behavioural flexibility evidence of cognitive complexity? How evolution can inform comparative cognition. *Interface Focus*, *7*(3), 20160121. https://doi.org/10.1098/rsfs.2016.0121
Post, W., Poston, J. P., & Bancroft, G. T. (1996). *Boat-tailed grackle: Quiscalus major*. American Ornithologists' Union.
R Core Team. (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org
Revelle, W. (2017). *Psych: Procedures for psychological, psychometric, and personality research*. Northwestern University. https://CRAN.R-project.org/package=psych

Rutz, C., Klump, B. C., Komarczyk, L., Leighton, R., Kramer, J., Wischnewski, S., Sugasawa, S., Morrissey, M. B., James, R., St Clair, J. J., et al. (2016). Discovery of species-wide tool use in the hawaiian crow. *Nature*, *537*(7620), 403–407.

Selander, R. K., & Giller, D. R. (1961). Analysis of sympatry of great-tailed and boat-tailed grackles. *The Condor*, *63*(1), 29–86.

Sevchik, A, Logan, CJ, Bergeron, L, Blackwell, A, Rowney, C, & Lukas, D. (2019). Investigating sex differences in genetic relatedness in great-tailed grackles in tempe, arizona to infer potential sex biases in dispersal. *In Principle Acceptance by PCI Ecology of the Version on 29 Nov 2019*. http://corinalogan.com/Preregistrations/gdispersal.html

Sol, D., Duncan, R. P., Blackburn, T. M., Cassey, P., & Lefebvre, L. (2005). Big brains, enhanced cognition, and response of birds to novel environments. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(15), 5460–5465. https://doi.org/10.1073/pnas.0408145102

Sol, D., & Lefebvre, L. (2000). Behavioural flexibility predicts invasion success in birds introduced to new zealand. *Oikos*, *90*(3), 599–605. https://doi.org/10.1034/j.1600-0706.2000.900317.x

Sol, D., Székely, T., Liker, A., & Lefebvre, L. (2007). Big-brained birds survive better in nature. *Proceedings of the Royal Society of London B: Biological Sciences*, *274*(1611), 763–769.

Sol, D., Timmermans, S., & Lefebvre, L. (2002). Behavioural flexibility and invasion success in birds. *Animal Behaviour*, *63*(3), 495–502.

Stan Development Team. (2020). *RStan: The R interface to Stan*. http://mc-stan.org/

Taylor, A. H., Hunt, G. R., Holzhaider, J. C., & Gray, R. D. (2007). Spontaneous metatool use by new caledonian crows. *Current Biology*, *17*(17), 1504–1507.

Team, S. D. et al. (2019). Stan modeling language users guide and reference manual. *Technical Report*.

Wehtje, W. (2003). The range expansion of the great-tailed grackle (quiscalus mexicanus gmelin) in north america since 1880. *Journal of Biogeography*, *30*(10), 1593–1607. https://doi.org/10.1046/j.1365-2699.2003.00970.x

Wiens, J. A. (1997). Metapopulation dynamics and landscape ecology. In *Metapopulation biology* (pp. 43–62). Elsevier.

Wright, T. F., Eberhard, J. R., Hobson, E. A., Avery, M. L., & Russello, M. A. (2010). Behavioral flexibility and species invasions: The adaptive flexibility hypothesis. *Ethology Ecology & Evolution*, *22*(4), 393–404.

Wu, J., Jenerette, G. D., Buyantuyev, A., & Redman, C. L. (2011). Quantifying spatiotemporal patterns of urbanization: The case of the two fastest growing metropolitan regions in the united states. *Ecological Complexity*, *8*(1), 1–8.