

Using repeatability of performance within and across contexts to validate measures of behavioral flexibility

McCune KB^{1*} Blaisdell AP² Johnson-Ulrich Z¹ Lukas D³
MacPherson M¹ Seitz B² Sevchik A⁴ Logan CJ³

2023-01-31

Open...  access  code  peer review  data

Affiliations: 1) University of California Santa Barbara, USA, 2) University of California Los Angeles, USA, 3) Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany, 4) Arizona State University, Tempe, AZ USA. *Corresponding author: kelseybmccune@gmail.com

This is a revision of the post-study manuscript of the preregistration that was pre-study peer reviewed and received an In Principle Recommendation on 26 Mar 2019 by:

Aur lie Coulon (2019) Can context changes improve behavioral flexibility? Towards a better understanding of species adaptability to environmental changes. *Peer Community in Ecology*, 100019. [10.24072/pci.ecology.100019](https://doi.org/10.24072/pci.ecology.100019). Reviewers: Maxime Dahirel and Andrea Griffin

Preregistration: [html](#), [pdf](#), [rmd](#)

Post-study manuscript we submitted the first version of the post-study manuscript to PCI Ecology for post-study peer review on 3 Jan 2022; we revised it per reviewer comments and this piece was split from the other, distinct components of the preregistrations and resubmitted on 15 Aug 2022; additional reviewer feedback is now incorporated and we resubmit this revised version to PCI Ecology: revised preprint [pdf](#) at EcoEvoRxiv, [rmd](#).

ABSTRACT

Research into animal cognitive abilities is increasing quickly and often uses methods where behavioral performance on a task is assumed to represent variation in the underlying cognitive trait. However, because these methods rely on behavioral responses as a proxy for cognitive ability, it is important to validate that the task structure does, in fact, target the cognitive trait of interest rather than non-target cognitive, personality, or motivational traits (construct validity). Although it can be difficult, or impossible, to definitively assign performance to one cognitive trait, one way to validate that task structure is more likely to elicit performance based on the target cognitive trait is to assess the temporal and contextual repeatability of performance. In other words, individual performance is likely to represent an inherent trait when it is consistent across time and across similar or different tasks that theoretically test the same trait. Here, we assessed the temporal and contextual repeatability of performance on tasks intended to test the cognitive trait behavioral flexibility in great-tailed grackles (*Quiscalus mexicanus*). For temporal repeatability, we quantified the number of trials to form a color preference after each of multiple color reversals on a serial reversal learning task. For

36 contextual repeatability, we then compared performance on the serial color reversal task to the latency to
37 switch among solutions on each of two different multi-access boxes. We found that the number of trials to
38 form a preference in reversal learning was repeatable across serial color reversals and the latency to switch
39 a preference was repeatable across color reversal learning and the multi-access box contexts. This supports
40 the idea that the reversal learning task structure elicits performance reflective of an inherent trait, and that
41 reversal learning and solution switching on multi-access boxes similarly reflect the inherent trait of behavioral
42 flexibility.

43 **KEYWORDS**

44 Behavioral flexibility, repeatability, construct validity, animal cognition

45 **INTRODUCTION**

46 Research on the cognitive abilities of non-human animals is important for several reasons. By understand-
47 ing animal cognitive abilities, we can clarify factors that influenced the evolution of human cognition, the
48 mechanisms that relate cognition to ecological and evolutionary dynamics, or we can use the knowledge to
49 facilitate more humane treatment of captive individuals (Shettleworth, 2010). In the last 50 years, compar-
50 ative psychologists and behavioral ecologists have led a surge in studies innovating methods for measuring
51 cognitive traits in animals. As a result, we have come to understand cognition as the process of acquiring
52 information, followed by storage, retrieval, and use of that information for guiding behavior (Shettleworth,
53 2010). Evidence now exists that various species possess cognitive abilities in both the physical (e.g. object
54 permanence: Salwiczek et al., 2009; causal understanding: Taylor et al., 2012) and social domains (e.g. social
55 learning: Hoppitt et al., 2012; transitive inference: MacLean et al., 2008).

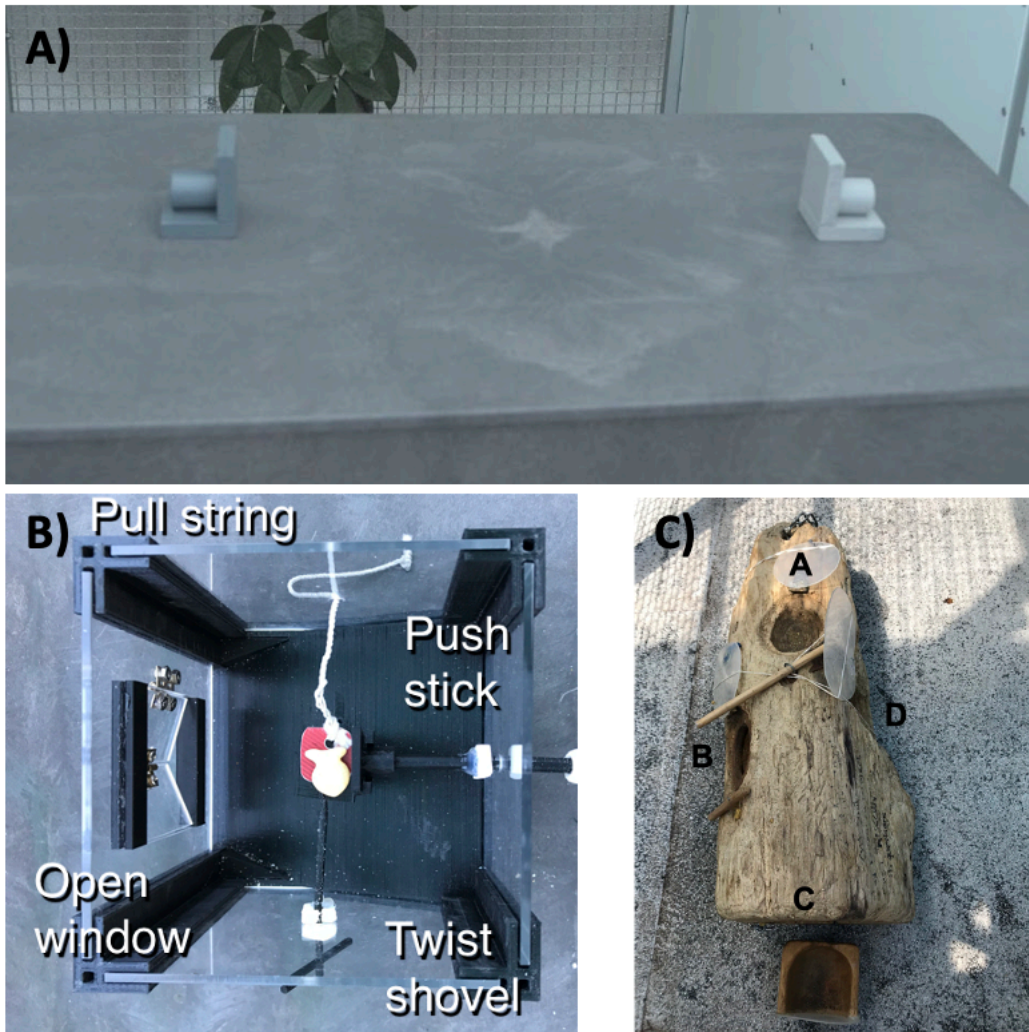
56 Cognitive traits are not directly observable and nearly all methods to quantify cognition use behavioral
57 performance as a proxy for cognitive ability. Consequently, it is important to evaluate the validity of the
58 chosen methods for quantifying a cognitive trait. To better understand whether performance on a type of
59 task is likely to reflect a target cognitive trait (i.e., that the method has construct validity), researchers can
60 test for repeatability in individual performance within and across tasks (Völter et al., 2018). However, while
61 many cognitive abilities have been tested, and various methods used, it is rare for one study to repeatedly test
62 individuals with the same method or use multiple methods to test for a given cognitive ability. This could
63 be problematic because cognitive traits are not directly observable, so nearly all methods use behavioral
64 performance as a proxy for cognitive ability. Using only one method to measure a cognitive trait could be
65 problematic because it is hard to discern whether non-target cognitive, personality, or motivational factors
66 may be the cause of variation in performance on the task (Morand-Ferron et al., 2016). For example, the
67 success of pheasants on multiple similar and different problem-solving tasks was related to individual variation
68 in persistence and motivation, rather than problem solving ability (Horik & Madden, 2016). Additionally,
69 performance on cognitive tasks can be affected by different learning styles, where individuals can vary
70 in their perception of the salience of stimuli within a task, the impact of a reward (or non-reward) on
71 future behavior, or the propensity to sample alternative stimuli (Rowe & Healy, 2014). By assessing the
72 temporal and contextual repeatability of performance, researchers can quantify the proportion of variation in
73 performance that is attributable to consistent individual differences likely to reflect the level of the cognitive
74 trait relative to other ephemeral factors that affect individual performance (Cauchoix et al., 2018).

75 Behavioral flexibility, the ability to change behavior when circumstances change, is a general cognitive ability
76 that likely affects interactions with both the social and physical environment (Bond et al., 2007). Although
77 by definition behavioral flexibility incorporates plasticity in behavior through learning, there is also evidence
78 that the ability to change behavior could be an inherent trait that varies among individuals and species. For
79 example, the pinyon jay - a highly social species of corvid - made fewer errors in a serial reversal learning
80 task than the more asocial Clark's nutcracker or Woodhouse's scrub-jay, but all three species exhibited
81 similar learning curves over successive reversals (Bond et al., 2007). This indicates that the three species
82 differed in the level of the inherent ability, but were similar in the plasticity of performance through learning.

83 Behavioral flexibility could be measured using a variety of methods (Mikhalevich et al., 2017), but the most
84 popular method is reversal learning (Bond et al., 2007) where behavioral flexibility is quantified as the speed
85 that individuals are able to switch a learned preference. However, to our knowledge, no studies have assessed
86 the construct validity of this task by comparing performance of individuals over time and across different
87 tasks that are predicted to require flexible behavior.

88 In the wild, this ability to change behavior when circumstances change is expected to result in individuals
89 and species that adapt quickly to novelty by showing a high rate of foraging innovations. For example,
90 cross-taxon correlational studies found that species that were “behaviorally flexible”, in that there were
91 many documented foraging innovations, were also more likely to become invasive when introduced to novel
92 habitats (Sol et al., 2002). The ability to innovate solutions to novel problems can also be more directly
93 quantified using a multi-access or puzzle box task, where the subject must use new behavior patterns to solve
94 the task to get food. While it is generally assumed that foraging innovation rate corresponds to the cognitive
95 ability behavioral flexibility (Sol et al., 2002), few studies compare innovation performance and solution
96 switching (a measure of flexibility) on a multi-access box task to performance on a different behavioral
97 flexibility task like reversal learning.

98 We tested two hypotheses about the construct validity of the reversal learning method as a measure of behav-
99 ioral flexibility in the great-tailed grackle (*Quiscalus mexicanus*; hereafter “grackle”). First, we determined
100 whether performance on a reversal learning task represents an inherent trait by assessing the repeatability of
101 performance across serial reversals (temporal repeatability). Secondly, we determined whether the inherent
102 trait measured by color reversal learning is likely to represent behavioral flexibility by assessing the cross-
103 contextual repeatability of performance on this task with another task also thought to measure flexibility.
104 Our previous research found that behavioral flexibility does affect innovation ability on a multi-access box
105 (C. Logan et al., 2022), so here our second hypothesis tested whether individuals show contextual repeata-
106 bility of flexibility by comparing performance on the color reversal learning task to the latency of solution
107 switching on two different multi-access boxes (Fig. 1). We chose solution switching because it requires
108 similar attention to changing reward contingencies, thus serving as a measure of flexibility, but in a different
109 context (e.g. the food is always visible, there is no color association learning required). In other words, in
110 both color reversal learning and solution switching individuals learned a preferred way to obtain food, but
111 then contingencies changed such that food can no longer be obtained with this learned preference and the
112 grackle must be able to switch to a new method. As a human-associated species, the grackle is an ideal
113 subject for this study because the rapid range expansion suggests that they adapted quickly in response to
114 human-induced rapid environmental change (Summers et al., 2022; Wehtje, 2003) and the genus *Quiscalus*
115 has a high rate of foraging innovations in the wild (Grabrucker & Grabrucker, 2010; Lefebvre & Sol, 2008).
116 Therefore, as their environment may select for flexible and innovative behavior, we believe that these tasks
117 are ecologically relevant and will elicit individual variation in performance.



118

119 **Figure 1.** We assessed flexibility as the latency to switch a preference across 3 contexts illustrated here. A)
 120 We used two colored containers (tubes) in a color reversal learning task, as well as B) plastic and C) wooden
 121 multi-access boxes that each had 4 possible ways (loci) to access food. In each context, after a preference
 122 for a color/locus was formed, we made the preferred choice non-functional and then measured the latency of
 123 the grackle to switch to a new color/locus.

124 METHODS

125 The hypotheses, methods, and analysis plan for this research are described in detail in the [peer-reviewed](#)
 126 [preregistration](#). We give a short summary of these methods here, with any changes from the preregistration
 127 summarized in the *Deviations from the preregistration* section below and further explained in the updates
 128 to the preregistration (indicated in italics).

129 Preregistration details

130 This experiment was one piece (**H3a and H3b**) of a larger project. This project is detailed in the prereg-
 131 istration that was written (2017), submitted to PCI Ecology for peer review (July 2018), and received the
 132 first round of peer reviews a few days before data collection began (Sep 2018). We revised and resubmitted

133 this preregistration after data collection had started (Feb 2019) and it passed peer review (Mar 2019) before
134 any of the planned analyses had been conducted. See the [peer review history](#) at PCI Ecology.

135 **Summary of hypotheses**

136 Our first hypothesis considered whether behavioral flexibility (as measured by reversal learning of a color
137 preference) would be repeatable within individuals across serial reversals. Secondly, we hypothesized that,
138 as an inherent trait, behavioral flexibility results in repeatable performance across other contexts (Fig. 1)
139 that require changing behavior when circumstances change (context 1=reversal learning on colored tubes,
140 context 2=plastic multi-access box, context 3=wooden multi-access box).

141 **Summary of methods**

142 **Subjects** Great-tailed grackles were caught in the wild in Tempe, Arizona USA using a variety of trapping
143 methods. All individuals received color leg bands for individual identification and some individuals (n=34)
144 were brought temporarily into aviaries. Grackles were individually housed in an aviary (each 244 cm long
145 by 122 cm wide by 213 cm tall) for a maximum of six months where they had *ad lib* access to water at all
146 times. During testing, we removed their maintenance diet for up to four hours per day. During this time,
147 they had the opportunity to receive high value food items by participating in tests. Individuals were given
148 three to four days to habituate to the aviaries before we began testing.

149 **Serial color reversal learning** We first used serial reversal learning to measure grackle behavioral flex-
150 ibility. Briefly, we trained grackles to search in one of two differently colored containers for food (Fig.
151 1A). We used a random number generator to select the color (e.g. light gray) of the container that would
152 consistently contain a food reward across the initial trials. Within each trial, grackles could choose only
153 one container to look in for food. Eventually, grackles showed a significant preference for the rewarded
154 color container (where preference was defined as a minimum of 17 out of 20 correct choices), completing the
155 initial discrimination trials. We then switched the location of the food to the container of the other color
156 (a reversal). The food reward was then consistently located in the container of this second color (e.g. dark
157 gray) across trials until the grackles learned to switch their preference, after which we would again reverse
158 the food to the original colored container (e.g. light gray) and so on back and forth until they passed the
159 serial reversal learning experiment passing criterion (formed a preference in 2 sequential reversals in 50 or
160 fewer trials: C. Logan et al., 2022). We measured behavioral flexibility on each reversal as the time it took
161 grackles to switch their preference and search in the second colored container on a minimum of 17 out of 20
162 trials. See the protocol for serial reversal learning [here](#).

163 **Multi-access boxes** We additionally used two different multi-access boxes (hereafter “MAB”) to assess
164 behavioral flexibility as the latency to switch loci when a preferred locus becomes non-functional. All grackles
165 were given time to habituate to the MABs prior to testing. We set up the MABs in the aviary of each grackle
166 with food in and around each apparatus in the days prior to testing. At this point all loci were absent or fixed
167 in open, non-functional positions to prevent any early learning of how to solve each apparatus. We began
168 testing when the grackle was eating comfortably from the MAB. For each MAB, the goal was to measure how
169 quickly the grackle could learn to solve each locus, and then how quickly they could switch to attempting to
170 solve a new locus. Consequently, we measured the number of trials to solve a locus and the number of trials
171 until the grackle attempted a new locus after a previously solved locus was made non-functional (solution
172 switching). See protocols for MAB habituation and testing [here](#).

173 **Plastic multi-access box** This apparatus consisted of a box with transparent plastic walls (Fig. 1B).
174 There was a pedestal within the box where the food was placed and 4 different options (loci) set within the
175 walls for accessing the food. One locus was a window that, when opened, allowed the grackle to reach in to
176 grab the food. The second locus was a shovel that the food was placed on such that, when turned, the food
177 fell from the pedestal and rolled out of the box. The third locus was a string attached to a tab that the food

178 was placed on such that, when pulled, the food fell from the pedestal and rolled out of the box. The last
179 locus was a horizontal stick that, when pushed, would shove the food off the pedestal such that it rolled out
180 of the box. Each trial was 10 minutes long, or until the grackle used a locus to retrieve the food item. We
181 reset the box out of view of the grackle to begin the next trial. To pass criterion for a locus, the grackle had
182 to get food out of the box after touching the locus only once (i.e. used a functional behavior to retrieve the
183 food) in more than 2 trials across 2 sessions. Afterward, the locus is made non-functional to encourage the
184 grackle to interact with the other loci.

185 **Wooden multi-access box** This apparatus consisted of a natural log that contained 4 compartments (loci)
186 covered by transparent plastic doors (Fig. 1C). Each door opened in a different way (open up like a hatch,
187 out to the side like a car door, pull out like a drawer, or push in). During testing, all doors were closed and
188 food was placed in each locus. Each trial lasted 10 minutes or until the grackle opened a door. After solving
189 a locus, the experimenter re-baited that compartment, closed the door out of view of the grackle, and the
190 next trial began. After a grackle solved one locus 3 times, that door was fixed in the open position and the
191 compartment left empty to encourage the grackle to attempt the other loci.

192 **Repeatability analysis** Repeatability is defined as the proportion of total variation in performance that is
193 attributable to differences among individuals (Nakagawa & Schielzeth, 2010). In other words, performance
194 is likely to represent an inherent trait, when variation in performance is greater among individuals than
195 within individuals.

196 To measure repeatability within an individual across serial reversals of a color preference, we modeled the
197 number of trials to pass a reversal (choosing correctly on at least 17 out of 20 sequential trials) as a function
198 of the reversal number (i.e., first time the rewarded color is reversed, second time, third time, etc.) and
199 a random effect for individual. The reversal number for each grackle ranged between 6 to 11 (mean =
200 7.6) reversals, and the range was based on when individuals were able to pass two sequential reversals in
201 50 or fewer trials, or (in 1 case) when we reached the maximum duration that we were permitted to keep
202 grackles in the aviaries and they needed to be released. The variance components for the random effect and
203 residual variance were then used to determine the proportion of variance attributable to differences among
204 individuals.

205 By design in the serial reversal learning experiment, to reach the experiment ending criteria grackles became
206 faster at switching across serial reversals. We did attempt to run a model that additionally included a
207 random slope to test whether there were consistent individual differences in the rate that grackles switched
208 their preferences across reversals. However, we could not get the model to converge with our sample size
209 and the uninformative priors that were preregistered. We felt most comfortable using the preregistered
210 methods to avoid biasing the model output. To determine the statistical significance of the repeatability
211 value, while accounting for this non-independence of a change in reversal speed over time, we compared the
212 actual performance on the number of trials to switch a preference in each reversal to simulated data where
213 birds performed randomly within each reversal.

214 We tested for contextual repeatability by modeling the variance in latency (in seconds) to switch a preference
215 among and within individuals across 3 behavior switching contexts. Note that the time it took to switch a
216 colored tube preference in serial reversal learning was measured in trials, but the time it took to switch loci in
217 the MAB experiment was measured in seconds. We used the trial start times in the serial reversal experiment
218 to convert the latency to switch a preference from number of trials to number of seconds. Therefore, the
219 contexts across which we measured repeatability of performance were the latency to switch a preference to
220 a new color in the color reversal learning task and latency to switch to a new locus after a previously solved
221 locus was made non-functional on both MABs.

222 We used the DHARMA package (Hartig, 2019) in R to test whether our model fit our data and was not
223 heteroscedastic, zero-inflated or over-dispersed. We used the MCMCglmm package (Hadfield, 2010), with
224 uninformative priors, to model the relationships of interest for our two hypotheses.

225 Open data

226 All data are available at the Knowledge Network for Biocomplexity's data repository: <https://knb.ecoinformatics.org/view/doi:10.5063/F1VX0F0W> (K. McCune et al., 2022).
227

228 Deviations from the preregistration

229 In the middle of data collection

230 1) We originally planned to use a touchscreen test of serial reversal learning as one of the contexts in
231 this experiment. However, on 10 April 2019 we **discontinued the reversal learning experiment**
232 **on the touchscreen** because it appears to measure something other than what we intended to test
233 and it requires a huge time investment for each bird (which consequently reduces the number of other
234 tests they are available to participate in). This is not necessarily surprising because this is the first
235 time touchscreen tests have been conducted in this species, and also the first time (to our knowledge)
236 this particular reversal experiment has been conducted on a touchscreen with birds. We based this
237 decision on data from four grackles (2 in the flexibility manipulation group and 2 in the flexibility
238 control group; 3 males and 1 female). All four of these individuals showed highly inconsistent learning
239 curves and required hundreds more trials to form each preference when compared to the performance
240 of these individuals on the colored tube reversal experiment. It appears that there is a confounding
241 variable with the touchscreen such that they are extremely slow to learn a preference as indicated
242 by passing our criterion of 17 correct trials out of the most recent 20. We will not include the data
243 from this experiment when conducting the cross-test comparisons in the Analysis Plan section of the
244 preregistration.

245 2) 16 April 2019: Because we discontinued the touchscreen reversal learning experiment, we **added an**
246 **additional but distinct multi-access box** task, which allowed us to continue to measure flexibility
247 across three different experiments. There are two main differences between the first multi-access box,
248 which is made of plastic, and the new multi-access box, which is made of wood. First, the wooden
249 multi-access box is a natural log in which we carved out 4 compartments. As a result, the apparatus and
250 solving options are more comparable to what grackles experience in the wild, though each compartment
251 is covered by a transparent plastic door that requires different behaviors to open. Furthermore, there
252 is only one food item available in the plastic multi-access box and the bird could use any of 4 loci
253 to reach it. In contrast, the wooden multi-access box has a piece of food in each of the 4 separate
254 compartments.

255 Post data collection, pre-data analysis

256 3) We completed our simulation to explore the lower boundary of a minimum sample size and determined
257 that **our sample size for the Arizona study site is above the minimum** (see details and code
258 in [Ability to detect actual effects](#); 17 April 2020).

259 4) We originally planned on testing only **adults** to have a better understanding of what the species is
260 capable of, assuming the abilities we are testing are at their optimal levels in adulthood, and so we
261 could increase our statistical power by eliminating the need to include age as an independent variable
262 in the models. Because the grackles in Arizona were extremely difficult to catch, we ended up testing
263 two juveniles in this experiment. The juveniles' performance on the three tests was similar to the
264 adults, therefore we decided not to add age as an independent variable in the models to avoid reducing
265 our statistical power.

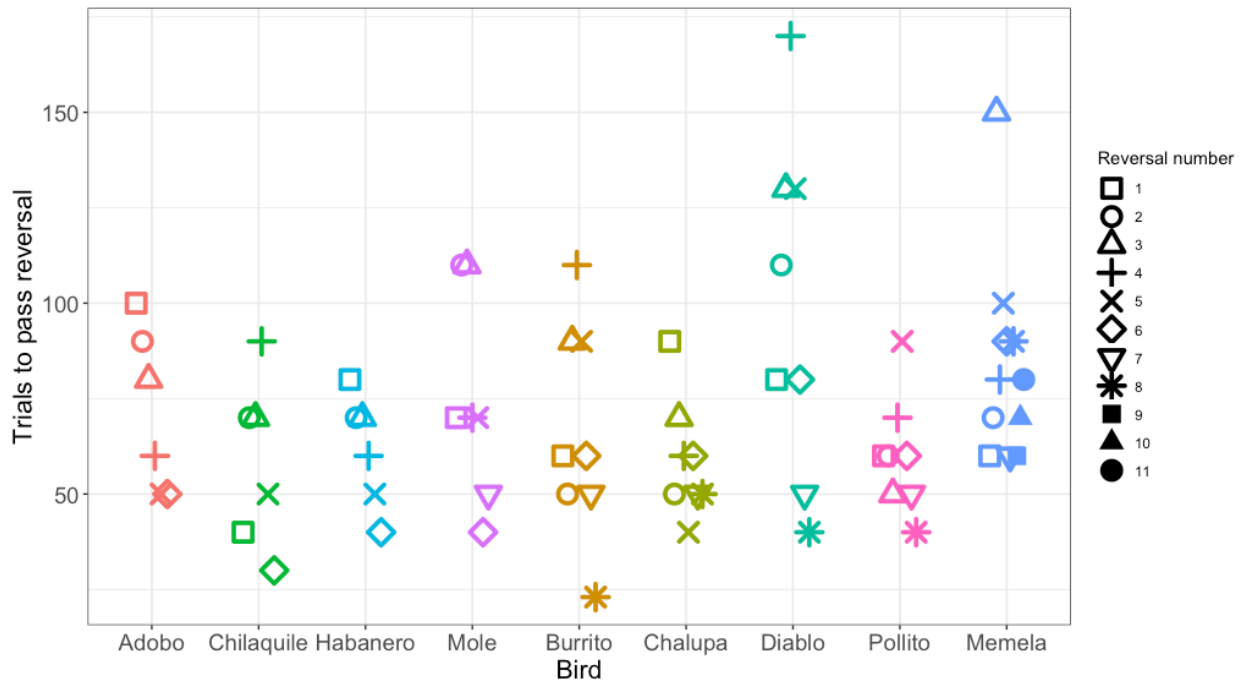
266 Post data collection, mid-data analysis

267 5) The distribution of values for the “number of trials to reverse” response variable in the **P3a analysis**
 268 was not a good fit for the Poisson distribution because it was overdispersed and heteroscedastic. We
 269 log-transformed the data to approximate a normal distribution and it passed all of the data checks.
 270 Therefore, we used a Gaussian distribution for our model, which fits the log-transformed data well.
 271 (24 Aug 2021)

272 6) We realized we mis-specified the model and variables for evaluating cross-contextual repeatability **P3b**
 273 **analysis**. The dependent variable should be latency to switch to a new preference (we previously
 274 listed “number of trials to solve”, which is more likely indicative of innovation rather than flexibility).
 275 Furthermore, to assess performance across contexts, this dependent variable should be the latency to
 276 switch in each of the 3 contexts. Note that the time it took to switch a colored tube preference in serial
 277 reversal learning was measured in trials, but the time it took to switch loci in the MAB experiment
 278 was measured in seconds. We used the trial start times in the serial reversal experiment to convert the
 279 latency to switch a preference from number of trials to number of seconds. In line with this change
 280 in the dependent variable, the independent variables are only Context (MAB plastic, MAB wood,
 281 reversal learning), and reversal number (the number of times individuals switched a preference when
 282 the previously preferred color/locus was made non-functional). Additionally, this dependent variable
 283 was heteroscedastic when we used a Poisson model, but passed all data checks when we log-transformed
 284 it to use a Gaussian model.

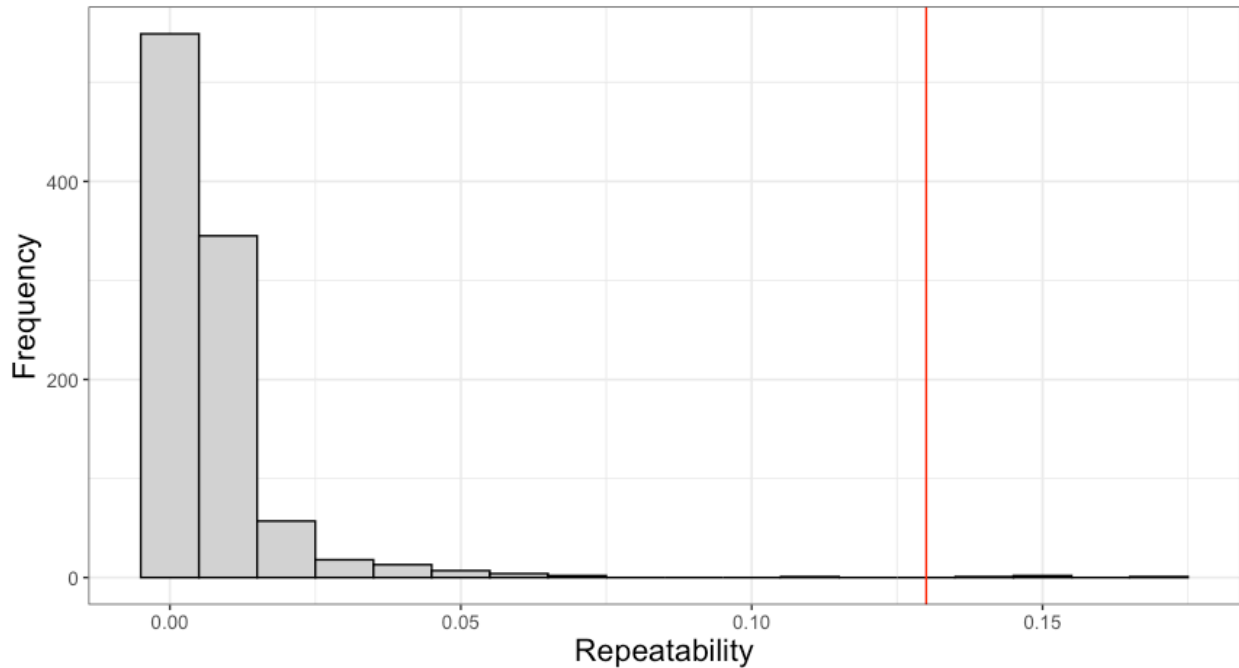
285 RESULTS

286 Our sample size was 9 for our first hypothesis testing temporal repeatability of reversal learning performance.
 287 Performance was repeatable within individuals within the context of reversal learning (Fig. 2): we obtained a
 288 repeatability value of 0.13 (95% credible interval (CI) = 4.64×10^{-16} - 0.43). We found that this repeatability
 289 value was significantly greater than expected if birds were performing randomly ($p=0.003$; Fig. 3; see
 290 analysis details in the R code for Analysis Plan > P3a). Consequently, and as preregistered, we did not
 291 need to conduct the analysis for the P3a alternative to determine whether a lack of repeatability was due to
 292 motivation or hunger.



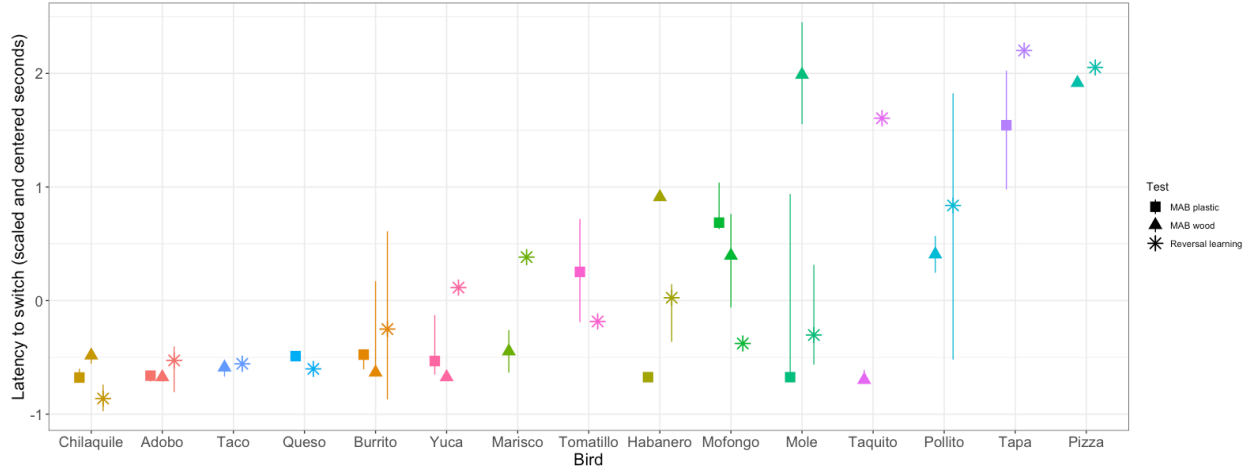
293 Figure 2: The number of trials each individual took to reverse a preference across serial reversals. The
 294 clustering of data points within each individual illustrates the temporal repeatability in performance. Each
 295

296 reversal is indicated by a different shape. Individuals are grouped by color and arranged from fastest to
 297 slowest to complete the serial reversal experiment. Note that as per the serial reversal experimental design,
 298 data from nearly all individuals include 2 reversals at or below 50 trials. The one exception was Memela,
 299 who never increased the speed to switch her preference.



300
 301 Figure 3: To determine the significance of our repeatability value while accounting for the non-independence
 302 of the serial reversal learning experimental design, we compared our repeatability value to repeatability
 303 values calculated from simulated data where birds performed randomly within each reversals. The red line
 304 indicates our observed value, and it is significantly larger than the repeatability values retrieved from the
 305 simulated data. This indicates that despite the design of the serial reversal learning experiment leading
 306 to a general increase in the speed that grackles pass each reversal, there were still consistent individual
 307 differences in performance across time.

308 We then assessed the repeatability of performance across contexts by quantifying whether individuals that
 309 were fast to switch a preference in the color reversal task were also fast to switch to attempting a new solution
 310 after passing criterion on a different solution on the two MAB tasks. We converted our metric of reversal
 311 speed from trials to reverse to seconds to reverse so the measures across contexts would be on the same scale.
 312 We had repeated measures across contexts for 15 grackles that participated in at least one color reversal
 313 and one solution switch on either or both MAB tasks. We found significant repeatability across contexts
 314 ($R=0.36$, $CI = 0.10 - 0.64$, $p=0.01$; Fig. 4), where latency to switch was consistent within individuals and
 315 different among individuals.



316

317 Figure 4: Grackle performance on the different contexts for measuring behavioral flexibility: multi-access
 318 box (MAB) plastic (square symbol), MAB wood (triangle symbol), and reversal learning with color tubes
 319 (star symbol). Points indicate the (scaled and centered) median performance of an individual in each
 320 context, the lines indicate the variation in performance across multiple switches within a context. Some
 321 individuals participated in a context, but did not experience multiple preference switches and so there is a
 322 point, but no line. Additionally, some individuals are missing points for a given context because they did
 323 not participate. Grackles are ordered on the x-axis from fastest (left) to slowest (right).

324 DISCUSSION

325 We found that individual grackles were consistent in their behavioral flexibility performance during multiple
 326 assessments within the same context, and across multiple assessments in different contexts. This indicates
 327 that 1) the different methods we used to measure behavioral flexibility all likely measure the same inherent
 328 trait and 2) there is consistent individual variation in behavioral flexibility, which could impact other traits
 329 such as survival and fitness in novel areas, foraging, or social behavior.

330 In behavioral and cognitive research on animals, it is important to determine that the chosen method mea-
 331 sures the trait of interest (construct validity). Many experimental methods may lack construct validity
 332 because they were adapted from research on other species (e.g. from humans: Wood et al., 1980), applied to
 333 new contexts (e.g. from captive to wild animals: K. B. McCune et al., 2019), or created from an anthropomor-
 334 phic perspective (e.g. mirror self recognition tasks: Kohda et al., 2022). Funding and logistical limitations
 335 result in few researchers assessing the appropriateness of their methods by testing construct validity through
 336 convergent (similar performance across similar tasks) and discriminant validity (different performance across
 337 different tasks). Although our sample size was small, which likely led to moderately large credible intervals,
 338 we still found significant temporal and contextual repeatability of switching performance. This evidence for
 339 convergent validity indicates these similar tasks are likely assessing the same latent trait of interest (Morand-
 340 Ferron et al., 2022; Völter et al., 2018). However, it is important to also test for discriminant validity by
 341 comparing performance on flexibility tasks with tasks intended to measure different cognitive abilities. For
 342 example, it is possible that performance on serial reversal learning and solution switching on the MAB tasks
 343 is reflective of a trait other than behavioral flexibility, like inhibition (MacLean et al., 2014). Indeed, we
 344 previously found that the more flexible grackles on the serial reversal learning task were also better able to
 345 inhibit responding to a non-rewarded stimulus in a go/no-go task to measure self-control (Logan et
 346 al., 2021). Consequently, more research is needed to interpret whether some aspect of performance on the
 347 go/no-go task reflects behavioral flexibility or whether performance on the reversal learning task is influenced
 348 by inhibition.

349 The functional importance of behavioral flexibility is that it is thought to facilitate invasion success by
 350 allowing individuals to quickly change their behavior when circumstances change. For example, flexible
 351 grackles may innovate new foraging techniques or generalize standard techniques to new food items in

352 novel areas. The great-tailed grackle has rapidly expanded its range (Summers et al., 2022; Wehtje, 2003),
353 implying that it is able to have high survival and fitness in the face of environmental change. Although
354 grackles are a behaviorally flexible species (Logan, 2016), we show here that there are consistent individual
355 differences among grackles in how quickly they are able to change their behavior when circumstances change
356 in multiple different contexts. While some grackles were consistently faster at changing their behavior
357 (e.g., Chilaquile), others were consistently slower (e.g., Yuca). This consistency in performance may seem
358 contradictory to our previous research where we found that we are able to manipulate grackles to be more
359 flexible using serial reversal learning (C. Logan et al., 2022). That behavioral flexibility is both repeatable
360 within individuals across reversals, indicating it is an inherent trait, as well as being manipulatable through
361 serial reversals, aligns with the idea of behavioral reaction norms (Sih, 2013). This idea states that individuals
362 can show consistent individual differences in the baseline or average values of a trait of interest across time
363 or contexts, but the plasticity in the expression of the trait can also consistently vary among individuals.
364 Due to our small sample size, we were not able to explicitly test for behavioral reaction norms, but this is
365 an important next step in understanding consistent individual variation in behavioral flexibility in relation
366 to rapid environmental change. Past experience (developmental or evolutionary) with environmental change
367 influences how plastic the individuals are able to be (Sih, 2013). To understand the implications of this
368 individual variation in performance in this species that has experienced much environmental change during
369 the range expansion, our future research investigates how behavioral flexibility may relate to proximity to
370 the range edge (Logan CJ et al., 2020), and the variety of foraging techniques used in the wild (Logan CJ
371 et al., 2019).

372 By first validating the experimental methods for behavioral and cognitive traits, such that we are more
373 certain that our tests are measuring the intended trait, we are better able to understand the causes and
374 consequences of species, population, and individual differences. Individual variation in behavioral flexibility
375 has the potential to influence species adaptation and persistence under human-induced rapid environmen-
376 tal change (Sih, 2013). Consequently, we believe the results presented here are a timely addition to the
377 field by demonstrating two potential methods for measuring behavioral flexibility that produced repeatable
378 performance in at least one system.

379 ETHICS

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

- 381 1) US Fish and Wildlife Service (scientific collecting permit number MB76700A-0,1,2)
- 382 2) US Geological Survey Bird Banding Laboratory (federal bird banding permit number 23872)
- 383 3) Arizona Game and Fish Department (scientific collecting license number SP594338 [2017], SP606267
384 [2018], and SP639866 [2019])
- 385 4) Institutional Animal Care and Use Committee at Arizona State University (protocol number 17-1594R)
- 386 5) University of Cambridge ethical review process (non-regulated use of animals in scientific procedures:
387 zoo4/17 [2017])

388 AUTHOR CONTRIBUTIONS

389 **McCune:** Added MAB log experiment, hypothesis development, protocol development, data collection,
390 data interpretation, write up, revising/editing, materials.

391 **Blaisdell:** Prediction revision, assisted with programming the reversal learning touchscreen experiment,
392 protocol development, data interpretation, revising/editing.

393 **Johnson-Ulrich:** Prediction revision, programming, data collection, data interpretation, revising/editing.

394 **Lukas:** Hypothesis development, simulation development, data interpretation, revising/editing.

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

396 **Seitz:** Prediction revision, programmed the reversal learning touchscreen experiment, protocol development,
397 data interpretation, revising/editing.

398 **Sevchik:** Data collection, revising/editing.

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

401 FUNDING

402 This research is funded by the Department of Human Behavior, Ecology and Culture at the Max Planck Insti-
403 tute for Evolutionary Anthropology (2017-current), and by a Leverhulme Early Career Research Fellowship
404 to Logan (2017-2018).

405 CONFLICT OF INTEREST DISCLOSURE

406 We, the authors, declare that we have no financial conflicts of interest with the content of this article.
407 CJ Logan is a Recommender and, until 2022, was on the Managing Board at PCI Ecology. D Lukas is a
408 Recommender at PCI Ecology.

409 ACKNOWLEDGEMENTS

410 We thank our PCI Ecology recommender, Aurelie Coulon, and reviewers, Maxime Dahirel and Andrea
411 Griffin, for their feedback on this preregistration; Kevin Langergraber for serving as our ASU IACUC PI;
412 Ben Trumble and Angela Bond for logistical support; Melissa Wilson for sponsoring our affiliations at
413 Arizona State University and lending lab equipment; Kristine Johnson for technical advice on great-tailed
414 grackles; Arizona State University School of Life Sciences Department Animal Care and Technologies for
415 providing space for our aviaries and for their excellent support of our daily activities; Julia Cissewski for
416 tirelessly solving problems involving financial transactions and contracts; Sophie Kaube for logistical support;
417 Richard McElreath for project support; Aaron Blackwell and Ken Kosik for being the UCSB sponsors of
418 the Cooperation Agreement with the Max Planck Institute for Evolutionary Anthropology; Tiana Lam,
419 Anja Becker, and Brynna Hood for interobserver reliability video coding; Sawyer Lung for field support;
420 Alexis Breen for coding multi-access box videos; and our research assistants: Aelin Mayer, Nancy Rodriguez,
421 Brianna Thomas, Aldora Messinger, Elysia Mamola, Michael Guillen, Rita Barakat, Adriana Boderash,
422 Olateju Ojekunle, August Sevchik, Justin Huynh, Jennifer Berens, Amanda Overholt, Michael Pickett, Sam
423 Munoz, Sam Bowser, Emily Blackwell, Kaylee Delcid, Sofija Savic, Brynna Hood, Sierra Planck, and Elise
424 Lange.

425 SUPPLEMENTARY MATERIALS

426 D. PREREGISTRATION (detailed methods)

427 HYPOTHESES

428 **H3a: Behavioral flexibility within a context is repeatable within individuals.** Repeatability of
429 behavioral flexibility is defined as the number of trials to reverse a color preference being strongly negatively
430 correlated within individuals with the number of reversals.

431 **P3a:** Individuals that are faster to reverse a color preference in the first reversal will also be faster to reverse
432 a color preference in the second, etc. reversal due to natural individual variation.

433 **P3a alternative:** There is no repeatability in behavioral flexibility within individuals, which could indicate
434 that performance is state dependent (e.g., it depends on their fluctuating motivation, hunger levels, etc.).
435 We will determine whether performance on colored tube reversal learning related to motivation by examining
436 whether the latency to make a choice influenced the results. We will also determine whether performance was
437 related to hunger levels by examining whether the number of minutes since the removal of their maintenance
438 diet from their aviary plus the number of food rewards they received since then influenced the results.

439 **H3b: The consistency of behavioral flexibility in individuals across contexts (context 1=re-**
440 **versal learning on colored tubes, context 2=multi-access boxes, context 3=reversal learning**
441 **on touchscreen) indicates their ability to generalize across contexts.** Individual consistency of
442 behavioral flexibility is defined as the number of trials to reverse a color preference being strongly positively
443 correlated within individuals with the latency to solve new loci on each of the multi-access boxes and with
444 the number of trials to reverse a color preference on a touchscreen (total number of touchscreen reversals =
445 5 per bird).

446 *If P3a is supported (repeatability of flexibility within individuals)...*

447 **P3b:** ...and flexibility is correlated across contexts, then the more flexible individuals are better at general-
448 izing across contexts.

449 **P3b alternative 1:** ...and flexibility is not correlated across contexts, then there is something that influences
450 an individual's ability to discount cues in a given context. This could be the individual's reinforcement history
451 (tested in P3a alternative), their reliance on particular learning strategies (one alternative is tested in H4),
452 or their motivation (tested in P3a alternative) to engage with a particular task (e.g., difficulty level of the
453 task).

454 **DEPENDENT VARIABLES** *P3a and P3a alternative 1*

455 Number of trials to reverse a preference. An individual is considered to have a preference if it chose the
456 rewarded option at least 17 out of the most recent 20 trials (with a minimum of 8 or 9 correct choices out
457 of 10 on the two most recent sets of 10 trials). We use a sliding window to look at the most recent 10 trials
458 for a bird, regardless of when the testing sessions occurred.

459 *P3b: additional analysis: individual consistency in flexibility across contexts + flexibility is correlated across*
460 *contexts*

461 Number of trials to solve a new locus on the multi-access boxes *NOTE: Jul 2022 we realized this variable is*
462 *more likely to represent innovation, and we mean to assess flexibility here. Therefore we changed this variable*
463 *to latency to attempt to switch a preference after the previously rewarded color/locus becomes non-functional.*

464 **INDEPENDENT VARIABLES** *P3a: repeatable within individuals within a context*

- 465 1) Reversal number
- 466 2) ID (random effect because repeated measures on the same individuals)

467 *P3a alternative 1: was the potential lack of repeatability on colored tube reversal learning due to motivation*
468 *or hunger?*

- 469 1) Trial number
- 470 2) Latency from the beginning of the trial to when they make a choice
- 471 3) Minutes since maintenance diet was removed from the aviary
- 472 4) Cumulative number of rewards from previous trials on that day

- 473 5) ID (random effect because repeated measures on the same individuals)
474 6) Batch (random effect because repeated measures on the same individuals). Note: batch is a test cohort,
475 consisting of 8 birds being tested simultaneously

476 *P3b: repeatable across contexts*

477 *NOTE: Jul 2022 we changed the dependent variable to reflect the general latency to switch a preference*
478 *(in any of the three tasks) and so IVs 3 (Latency to solve a new locus) & 4 (Number of trials to reverse*
479 *a preference), below, are redundant. Furthermore, we did not include the touchscreen experiment in this*
480 *manuscript (previously accounted for with IV 5; see the Deviations section). Therefore, despite being listed*
481 *here in the preregistration as IVs that we proposed to include in the P3b model, in our post-study manuscript*
482 *we did not include these IVs in the final model. The IVs instead consisted of: Reversal (switch) number,*
483 *Context (colored tubes, plastic multi-access box, wooden multi-access box) and ID (random effect because*
484 *there were repeated measures on the same individuals).*

- 485 1) Reversal (switch) number
486 2) Context (colored tubes, plastic multi-access box, wooden multi-access box, touchscreen)
487 3) Latency to solve a new locus
488 4) Number of trials to reverse a preference (colored tubes)
489 5) Number of trials to reverse a preference (touchscreen)
490 6) ID (random effect because repeated measures on the same individuals)

491 **ANALYSIS PLAN** *P3a: repeatable within individuals within a context (reversal learning)*

492 **Analysis:** Is reversal learning (colored tubes) repeatable within individuals within a context (reversal
493 learning)? We will obtain repeatability estimates that account for the observed and latent scales, and
494 then compare them with the raw repeatability estimate from the null model. The repeatability estimate
495 indicates how much of the total variance, after accounting for fixed and random effects, is explained by
496 individual differences (ID). We will run this GLMM using the MCMCglmm function in the MCMCglmm
497 package (Hadfield, 2010) with a Poisson distribution and log link using 13,000 iterations with a thinning
498 interval of 10, a burnin of 3,000, and minimal priors [V=1, nu=0; Hadfield (2014)]. We will ensure the
499 GLMM shows acceptable convergence [i.e., lag time autocorrelation values <0.01; Hadfield (2010)], and
500 adjust parameters if necessary.

501 **NOTE (Aug 2021):** our data checking process showed that the distribution of values of the data (number of
502 trials to reverse) in this model was not a good fit for the Poisson distribution because it was overdispersed
503 and heteroscedastic. However, when log-transformed the data approximate a normal distribution and pass
504 all of the data checks, therefore we used a Gaussian distribution for our model, which fits the log-transformed
505 data well.

506 To roughly estimate our ability to detect actual effects (because these power analyses are designed for
507 frequentist statistics, not Bayesian statistics), we ran a power analysis in G*Power with the following settings:
508 test family=F tests, statistical test=linear multiple regression: Fixed model (R² deviation from zero), type
509 of power analysis=a priori, alpha error probability=0.05. The number of predictor variables was restricted
510 to only the fixed effects because this test was not designed for mixed models. We reduced the power to 0.70
511 and increased the effect size until the total sample size in the output matched our projected sample size
512 (n=32). The protocol of the power analysis is here:

513 *Input:*

514 Effect size $f^2 = 0.21$

515 err prob = 0.05

516 Power (1- err prob) = 0.7

517 Number of predictors = 1

518 *Output:*

519 Noncentrality parameter = 6.7200000

520 Critical F = 4.1708768

521 Numerator df = 1

522 Denominator df = 30

523 Total sample size = 32

524 Actual power = 0.7083763

525 This means that, with our sample size of 32, we have a 71% chance of detecting a medium effect (approximated
526 at $f^2=0.15$ by Cohen, 1988).

527 *P3a alternative: was the potential lack of repeatability on colored tube reversal learning due to motivation or*
528 *hunger?*

529 **Analysis:** Because the independent variables could influence each other or measure the same variable, I will
530 analyze them in a single model: Generalized Linear Mixed Model [GLMM; MCMCglmm function, MCM-
531 Cglmm package; Hadfield (2010)] with a binomial distribution (called categorical in MCMCglmm) and logit
532 link using 13,000 iterations with a thinning interval of 10, a burnin of 3,000, and minimal priors ($V=1$, $\nu=0$)
533 (Hadfield, 2014). We will ensure the GLMM shows acceptable convergence [lag time autocorrelation values
534 <0.01 ; Hadfield (2010)], and adjust parameters if necessary. The contribution of each independent variable
535 will be evaluated using the Estimate in the full model. NOTE (Apr 2021): This analysis is restricted to data
536 from their first reversal because this is the only reversal data that is comparable across the manipulated and
537 control groups.

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

545 *Input:*

546 Effect size $f^2 = 0.31$

547 err prob = 0.05

548 Power (1- err prob) = 0.7

549 Number of predictors = 4

550 *Output:*

551 Noncentrality parameter = 11.4700000

552 Critical F = 2.6684369

553 Numerator df = 4

554 Denominator df = 32

555 Total sample size = 37

556 Actual power = 0.7113216

557 This means that, with our sample size of 32, we have a 71% chance of detecting a large effect (approximated
558 at $f^2=0.35$ by Cohen, 1988).

559 *P3b: individual consistency across contexts*

560 **Analysis:** Do those individuals that are faster to reverse a color preference also have lower latencies to switch
561 to new options on the multi-access box? A Generalized Linear Mixed Model [GLMM; MCMCglmm function,
562 MCMCglmm package; (Hadfield, 2010) will be used with a Poisson distribution and log link using 13,000
563 iterations with a thinning interval of 10, a burnin of 3,000, and minimal priors (V=1, nu=0) (Hadfield, 2014).
564 We will ensure the GLMM shows acceptable convergence [lag time autocorrelation values <0.01; Hadfield
565 (2010)], and adjust parameters if necessary. We will determine whether an independent variable had an
566 effect or not using the Estimate in the full model.

567 To roughly estimate our ability to detect actual effects (because these power analyses are designed for
568 frequentist statistics, not Bayesian statistics), we ran a power analysis in G*Power with the following settings:
569 test family=F tests, statistical test=linear multiple regression: Fixed model (R² deviation from zero), type
570 of power analysis=a priori, alpha error probability=0.05. We reduced the power to 0.70 and increased the
571 effect size until the total sample size in the output matched our projected sample size (n=32). The number
572 of predictor variables was restricted to only the fixed effects because this test was not designed for mixed
573 models. The protocol of the power analysis is here:

574 *Input:*

575 Effect size $f^2 = 0.21$

576 err prob = 0.05

577 Power (1- err prob) = 0.7

578 Number of predictors = 1

579 *Output:*

580 Noncentrality parameter = 6.7200000

581 Critical F = 4.1708768

582 Numerator df = 1

583 Denominator df = 30

584 Total sample size = 32

585 Actual power = 0.7083763

586 This means that, with our sample size of 32, we have a 71% chance of detecting a medium effect (approximated
587 at $f^2=0.15$ by Cohen, 1988).

588 REFERENCES

- 589 Bond, A. B., Kamil, A. C., & Balda, R. P. (2007). Serial reversal learning and the evolution of behavioral
590 flexibility in three species of north american corvids (gymnorhinus cyanocephalus, nucifraga columbiana,
591 aphelecoma californica). *Journal of Comparative Psychology*, 121(4), 372.
- 592 Cauchoix, M., Chow, P., Van Horik, J., Atance, C., Barbeau, E., Barragan-Jason, G., Bize, P., Boussard, A.,
593 Buechel, S. D., Cabirol, A., et al. (2018). The repeatability of cognitive performance: A meta-analysis.
594 *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1756), 20170281.
- 595 Cohen, J. (1988). *Statistical power analysis for the behavioral sciences 2nd edn.* Erlbaum Associates,
596 Hillsdale.
- 597 Grabrucker, S., & Grabrucker, A. M. (2010). Rare feeding behavior of great-tailed grackles (quiscalus
598 mexicanus) in the extreme habitat of death valley. *The Open Ornithology Journal*, 3(1).
- 599 Hadfield, J. (2010). MCMC methods for multi-response generalized linear mixed models: The MCMCglmm
600 r package. *Journal of Statistical Software*, 33(2), 1–22. <https://doi.org/10.18637/jss.v033.i02>

- 601 Hadfield, J. (2014). *MCMCglmm course notes*. [http://cran.r-project.org/web/packages/MCMCglmm/
602 vignettes/CourseNotes.pdf](http://cran.r-project.org/web/packages/MCMCglmm/vignettes/CourseNotes.pdf)
- 603 Hartig, F. (2019). *DHARMA: Residual diagnostics for hierarchical (multi-level / mixed) regression models*.
604 <http://florianhartig.github.io/DHARMA/>
- 605 Hoppitt, W., Samson, J., Laland, K. N., & Thornton, A. (2012). *Identification of learning mechanisms in a
606 wild meerkat population*.
- 607 Horik, J. O. van, & Madden, J. R. (2016). A problem with problem solving: Motivational traits, but not
608 cognition, predict success on novel operant foraging tasks. *Animal Behaviour*, *114*, 189–198.
- 609 Kohda, M., Sogawa, S., Jordan, A. L., Kubo, N., Awata, S., Satoh, S., Kobayashi, T., Fujita, A., & Bshary,
610 R. (2022). Further evidence for the capacity of mirror self-recognition in cleaner fish and the significance
611 of ecologically relevant marks. *PLoS Biology*, *20*(2), e3001529.
- 612 Lefebvre, L., & Sol, D. (2008). Brains, lifestyles and cognition: Are there general trends? *Brain, Behavior
613 and Evolution*, *72*(2), 135–144.
- 614 Logan, C. J. (2016). Behavioral flexibility and problem solving in an invasive bird. *PeerJ*, *4*, e1975. [https:
615 //doi.org/10.7717/peerj.1975](https://doi.org/10.7717/peerj.1975)
- 616 Logan, C. J., McCune, K., MacPherson, M., Johnson-Ulrich, Z., Rowney, C., Seitz, B., Blaisdell, A., Deffner,
617 D., & Wascher, C. (2021). *Are the more flexible great-tailed grackles also better at behavioral inhibition?*
618 <https://doi.org/10.31234/osf.io/vpc39>
- 619 Logan, C., Blaisdell, A., Johnson-Ulrich, Z., Lukas, D., MacPherson, M., Seitz, B., Sevchik, A., & McCune,
620 K. (2022). Behavioral flexibility is manipulatable and it improves flexibility and problem solving in a
621 new context. *EcoEvoRxiv*. <https://doi.org/https://doi.org/10.32942/osf.io/5z8xs>
- 622 Logan, CJ, Lukas D, Bergeron L, Folsom M, & McCune, K. (2019). Is behavioral flexibility related to
623 foraging and social behavior in a rapidly expanding species? *In Principle Acceptance by PCI Ecology of
624 the Version on 6 Aug 2019*. http://corinalogan.com/Preregistrations/g_flexforaging.html
- 625 Logan, CJ, McCune, KB, Chen, N, & Lukas, D. (2020). Implementing a rapid geographic range expansion
626 - the role of behavior and habitat changes. *In Principle Acceptance by PCI Ecology of the Version on 6
627 Oct 2020*. <http://corinalogan.com/Preregistrations/gxpopbehaviorhabitat.html>
- 628 MacLean, E. L., Hare, B., Nunn, C. L., Addessi, E., Amici, F., Anderson, R. C., Aureli, F., Baker, J. M.,
629 Bania, A. E., Barnard, A. M., et al. (2014). The evolution of self-control. *Proceedings of the National
630 Academy of Sciences*, *111*(20), E2140–E2148.
- 631 MacLean, E. L., Merritt, D. J., & Brannon, E. M. (2008). Social complexity predicts transitive reasoning in
632 prosimian primates. *Animal Behaviour*, *76*(2), 479–486.
- 633 McCune, K. B., Jablonski, P., Lee, S., & Ha, R. R. (2019). Captive jays exhibit reduced problem-solving
634 performance compared to wild conspecifics. *Royal Society Open Science*, *6*(1), 181311.
- 635 McCune, K., Blaisdell, A., Johnson-Ulrich, Z., Lukas, D., MacPherson, M., Seitz, B., Sevchik, A., & Lo-
636 gan, C. (2022). Using repeatability of performance within and across contexts to validate measures of
637 behavioral flexibility. *Knowledge Network for Biocomplexity, Data package*. [https://doi.org/10.5063/
638 F18K77JH](https://doi.org/10.5063/F18K77JH)
- 639 Mikhalevich, I., Powell, R., & Logan, C. (2017). Is behavioural flexibility evidence of cognitive complexity?
640 How evolution can inform comparative cognition. *Interface Focus*, *7*(3), 20160121. [https://doi.org/10.
641 1098/rsfs.2016.0121](https://doi.org/10.1098/rsfs.2016.0121)
- 642 Morand-Ferron, J., Cole, E. F., & Quinn, J. L. (2016). Studying the evolutionary ecology of cognition in
643 the wild: A review of practical and conceptual challenges. *Biological Reviews*, *91*(2), 367–389.
- 644 Morand-Ferron, J., Reichert, M. S., & Quinn, J. L. (2022). Cognitive flexibility in the wild: Individual dif-
645 ferences in reversal learning are explained primarily by proactive interference, not by sampling strategies,
646 in two passerine bird species. *Learning & Behavior*, *50*(1), 153–166.
- 647 Nakagawa, S., & Schielzeth, H. (2010). Repeatability for gaussian and non-gaussian data: A practical guide
648 for biologists. *Biological Reviews*, *85*(4), 935–956.
- 649 Rowe, C., & Healy, S. D. (2014). Measuring variation in cognition. *Behavioral Ecology*, *25*(6), 1287–1292.
- 650 Salwiczek, L. H., Emery, N. J., Schlinger, B., & Clayton, N. S. (2009). The development of caching and
651 object permanence in western scrub-jays (*aphelocoma californica*): Which emerges first? *Journal of
652 Comparative Psychology*, *123*(3), 295.
- 653 Shettleworth, S. J. (2010). *Cognition, evolution, and behavior*. Oxford university press.
- 654 Sih, A. (2013). Understanding variation in behavioural responses to human-induced rapid environmental

- 655 change: A conceptual overview. *Animal Behaviour*, 85(5), 1077–1088.
- 656 Sol, D., Timmermans, S., & Lefebvre, L. (2002). Behavioural flexibility and invasion success in birds. *Animal*
657 *Behaviour*, 63(3), 495–502.
- 658 Summers, J., Lukas, D., Logan, C., & Chen, N. (2022). The role of climate change and niche shifts in
659 divergent range dynamics of a sister-species pair. *EcoEvoRxiv*. <https://doi.org/10.32942/osf.io/879pe>
- 660 Taylor, A. H., Knaebe, B., & Gray, R. D. (2012). An end to insight? New caledonian crows can spontaneously
661 solve problems without planning their actions. *Proceedings of the Royal Society B: Biological Sciences*,
662 279(1749), 4977–4981.
- 663 Völter, C. J., Tinklenberg, B., Call, J., & Seed, A. M. (2018). Comparative psychometrics: Establishing
664 what differs is central to understanding what evolves. *Philosophical Transactions of the Royal Society B:*
665 *Biological Sciences*, 373(1756), 20170283.
- 666 Wehtje, W. (2003). The range expansion of the great-tailed grackle (*quiscalus mexicanus gmelin*) in north
667 america since 1880. *Journal of Biogeography*, 30(10), 1593–1607. [https://doi.org/10.1046/j.1365-2699.](https://doi.org/10.1046/j.1365-2699.2003.00970.x)
668 [2003.00970.x](https://doi.org/10.1046/j.1365-2699.2003.00970.x)
- 669 Wood, S., Moriarty, K. M., Gardner, B. T., & Gardner, R. A. (1980). Object permanence in child and
670 chimpanzee. *Animal Learning & Behavior*, 8(1), 3–9.