

1 Behavioral flexibility is manipulable and it improves flexibility and  
2 innovativeness in a new context.

3 Logan CJ<sup>1\*</sup>      Lukas D<sup>1\*</sup>      Blaisdell AP<sup>2</sup>      Johnson-Ulrich Z<sup>3</sup>      MacPherson M<sup>3</sup>  
4                      Seitz B<sup>2</sup>                      Sevchik A<sup>4</sup>                      McCune KB<sup>3</sup>

5 2023-04-17

6  Open...  access   code  data  peer review

7  
8 **Affiliations:** 1) Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany, 2) University of  
9 California Los Angeles, USA, 3) University of California Santa Barbara, USA, 4) Arizona State University,  
10 Tempe, AZ USA. \*Corresponding author: [corina\\_logan@eva.mpg.de](mailto:corina_logan@eva.mpg.de)

11  
12 **This is the post-study manuscript of the preregistration that was pre-study peer reviewed and**  
13 **received an In Principle Recommendation on 26 Mar 2019 by:**

14 Aurélie Coulon (2019) Can context changes improve behavioral flexibility? Towards a better un-  
15 derstanding of species adaptability to environmental changes. *Peer Community in Ecology*, 100019.  
16 [10.24072/pci.ecology.100019](https://doi.org/10.24072/pci.ecology.100019). Reviewers: Maxime Dahirel and Andrea Griffin

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

18 **Post-study manuscript** (submitted to PCI Ecology for post-study peer review on 3 Jan 2022, underwent  
19 3 rounds of revisions, R2 submitted Mar 2023, R3 submitted Apr 2023): preprint [pdf](#) at EcoEvoRxiv, [html](#),  
20 [rmd](#)

## 21 ABSTRACT

22 Behavioral flexibility, the ability to adapt behavior to new circumstances, is thought to play an important  
23 role in a species' ability to successfully adapt to new environments and expand its geographic range. However,  
24 flexibility is rarely directly tested in a way that would allow us to determine how flexibility works to predict  
25 a species' ability to adapt their behavior to new environments. We use great-tailed grackles (*Quiscalus*  
26 *mexicanus*; a bird species) as a model to investigate this question because they have recently rapidly expanded  
27 their range into North America. We attempted to manipulate grackle flexibility using shaded (light and dark  
28 gray) tube reversal learning to determine whether flexibility is generalizable across contexts (multi-access  
29 box), and what learning strategies grackles employ. We found that flexibility was manipulable: birds in the  
30 manipulated group took fewer trials to pass criterion with increasing reversal number, and they reversed a  
31 shade preference in fewer trials by the end of their serial reversals compared to control birds who had only  
32 one reversal. Birds that passed their last reversal faster were also more flexible (faster to switch between loci)  
33 and innovative (solved more loci) on a multi-access box. All grackles in the manipulated reversal learning  
34 group used one learning strategy (epsilon-decreasing) in all reversals, and none used a particular exploration  
35 or exploitation strategy earlier or later in their serial reversals. Understanding how flexibility causally relates

36 to other traits will allow researchers to develop robust theory about what flexibility is and when to invoke  
37 it as a primary driver in a given context, such as a rapid geographic range expansion.

## 38 [Video summary](#)

## 39 INTRODUCTION

40 Behavioral flexibility, the ability to adapt behavior to new circumstances through packaging information and  
41 making it available to other cognitive processes (see Mikhalevich et al., 2017 for the theoretical background  
42 on this definition), is thought to play an important role in a species' ability to successfully adapt to new  
43 environments and expand its geographic range (e.g., Lefebvre et al., 1997; Sol et al., 2002, 2005, 2007; Sol  
44 & Lefebvre, 2000). The behavioral flexibility (hereafter referred to as flexibility) of individuals is considered  
45 an important trait that facilitates the capacity for learning, which is then associated with problem solving  
46 ability (applying what one has learned about the world to then attempt to access a resource that is not  
47 readily accessible) (see review in Lea et al., 2020). It is hypothesized that, through flexibility, individuals  
48 can increase the diversity of their behaviors either via asocial learning (innovativeness) or social learning,  
49 leading to the establishment of the population in a new area (Wright et al., 2010).

50 It is predicted that flexibility should positively relate with innovativeness, the ability to create a new behavior  
51 or use an existing behavior in a new situation (Griffin & Guez, 2014). However, these predictions are based  
52 on species-level data and proxies for flexibility and for innovation (e.g., brain size, number of anecdotal  
53 reports of “novel” foods consumed) when examining such relationships (see Logan et al., 2018). Flexibility is  
54 rarely directly tested in species that are rapidly expanding their geographic ranges in a way that would allow  
55 us to determine how flexibility works and predict a species' ability to adapt their behavior to new areas.  
56 Those investigations that examine the relationship between flexibility and innovation or problem solving in  
57 species that are expanding their range show mixed results, with these variables correlating positively (e.g.,  
58 grey squirrels: Chow et al., 2016), negatively (e.g., Indian mynas: Griffin et al., 2013), or not at all (e.g.,  
59 stick tool use and string pulling in great-tailed grackles: Logan, 2016). Problem solving in these contexts  
60 involves experimental assays that do not necessarily require innovativeness to solve (e.g., the ability to solve  
61 tasks using pre-trained behaviors: Griffin & Guez, 2014). However, none of these experiments manipulated  
62 flexibility.

63 Here, we take the first step to improving our understanding of whether and how flexibility relates to innova-  
64 tiveness by starting with one population and performing a manipulative experiment on one of the variables to  
65 determine whether there is an associated change in the other. Once this association is known, future research  
66 can then investigate whether flexibility and innovativeness are involved in a range expansion. Manipulative  
67 experiments go beyond correlations to infer a cause and effect relationship between the manipulated variable  
68 and the variable(s) measured after the manipulation (Hernán & Robins, 2006; McElreath, 2020). A ma-  
69 nipulative experiment combined with the random assignment of subjects to a condition (manipulated group  
70 or control group), eliminates many confounds associated with internal and external variation (for example,  
71 season, motivation, sex, and so on). Such manipulative experiments in behavioral ecology have primarily  
72 been conducted in laboratory settings because of the increased feasibility, however such experiments are now  
73 also being conducted in wild settings (e.g., Aplin et al., 2015).

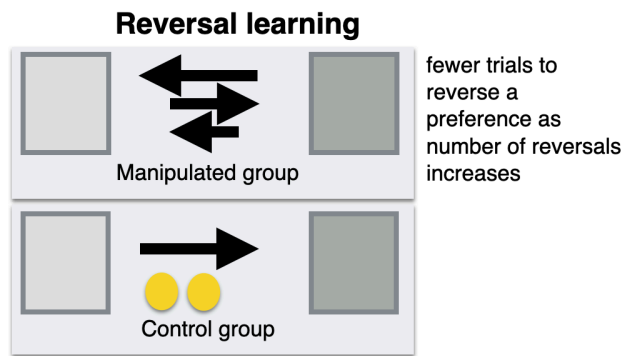
74 We focused our study on one population of great-tailed grackles (*Quiscalus mexicanus*, hereafter grackles),  
75 a bird species that is flexible (Logan, 2016). While they are originally from Central America, grackles  
76 have rapidly expanded their geographic range across the US since 1880 (Summers et al., 2023; Wehtje,  
77 2003). We attempted to manipulate grackle flexibility using serial reversals of a shade (light or dark gray)  
78 preference to determine whether their flexibility is generalizable across additional experimental contexts  
79 (touchscreen reversal learning and multi-access box solution switching), whether improving flexibility also  
80 improves innovativeness (number of loci solved on a multi-access box), and what learning strategies grackles  
81 employ (Figure 1).

82 Reversal learning is a common way of measuring flexibility that has been used for many decades across  
83 many species, therefore lending itself well to comparative analyses and generalizations (see review in Lea

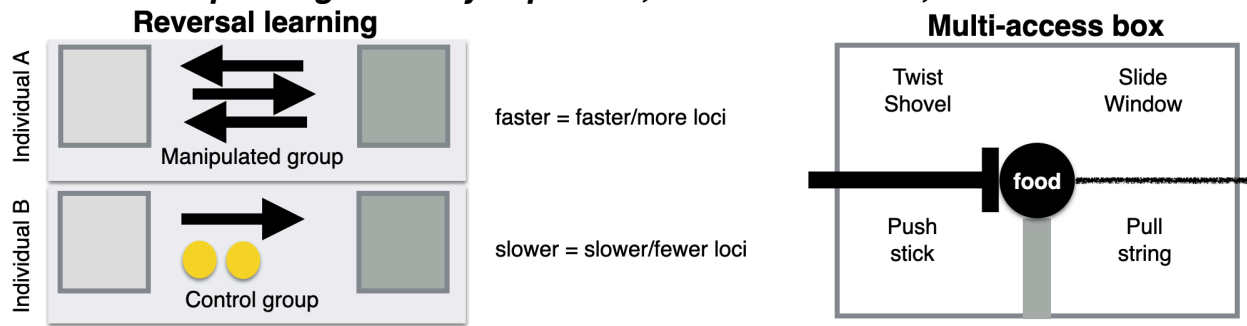
84 et al., 2020). In this test, an individual learns to prefer the rewarded option, which differs from the non-  
 85 rewarded option in shade/color, shape, space, or another discriminable feature. Once this initial preference is  
 86 formed, the previously non-rewarded option becomes the rewarded option and vice versa, and the preference  
 87 is reversed. Individuals who are faster to reverse their preference are considered more flexible - better able to  
 88 change their behavior when the circumstances change. Serial reversal learning involves continuing to reverse  
 89 the preference back and forth to determine whether individuals learn a “win-stay, lose-shift” rule that, when  
 90 the reward no longer follows the expected option, they should switch to preferring the other option (Spence,  
 91 1936; J. Warren, 1965; J. M. Warren, 1965). Once this rule is learned, it can then be applied to new contexts  
 92 and result in improved performance over individuals who have not learned this rule (J. M. Warren, 1965).  
 93 We randomly assigned individuals to a manipulated or control condition and used serial reversals (for the  
 94 manipulated group) to attempt to manipulate flexibility and determine whether the manipulated individuals  
 95 were then more flexible and more innovative in other contexts.

96 If grackle flexibility is manipulable using serial reversals, this would provide us with a useful tool for investi-  
 97 gating the relationship between flexibility and any number of other variables implicated in geographic range  
 98 expansions. It would provide researchers with a way to examine the direct links between, for example, flexi-  
 99 bility and exploration, to determine whether they are connected and in which direction, which could provide  
 100 insights into how populations establish in a new location if cross-population manipulations were conducted.  
 101 If the flexibility manipulation is not successful, this could indicate either that we did not manipulate the  
 102 right aspect of flexibility (e.g., perhaps training them to solve a variety of different types of tasks quickly  
 103 would be more effective) or that grackle flexibility is not a trait that is trainable.

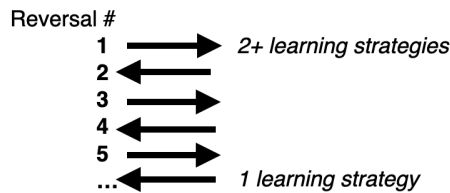
**A. Is flexibility manipulable?**



**B. Does manipulating flexibility improve it, & innovativeness, in a new context?**



**C. Do individuals converge on one learning strategy?**



104

105 **Figure 1.** A visual illustration of Hypothesis 1 (A), Hypothesis 2 (B), and Hypothesis 4 (C). Longer black  
 106 arrows indicate slower reversal times, the two yellow circles represent experience with the two yellow tubes  
 107 that both contained food for the control group.

108 PREREGISTERED HYPOTHESES

109 **H1: Behavioral flexibility, as measured by reversal learning using colored tubes, is manipulable.**

- 110 • **Prediction 1:** Individuals improve their flexibility on a serial reversal learning task using shaded  
111 tubes by generally requiring fewer trials to reverse a preference as the number of reversals increases  
112 (manipulation condition). Their flexibility on this test is manipulated relative to control birds who do  
113 not undergo serial reversals. Instead, individuals in the control condition are matched to manipulated  
114 birds for experience (they experience a similar number of trials), but there is no possibility of a  
115 functional tube preference because both tubes are the same shade (yellow) and both contain food,  
116 therefore either choice is correct.
- 117 • **P1 alternative 1:** If the number of trials to reverse a preference does not correlate with or positively  
118 correlates with reversal number, which would account for all potential correlation outcomes, this sug-  
119 gests that some individuals may prefer to rely on information acquired previously (i.e., they are slow  
120 to reverse) rather than relying on current cues (e.g., the food is in a new location) (Griffin & Guez,  
121 2014; Liu et al., 2016; e.g., Manrique et al., 2013; but see Homberg et al., 2007).

122 **H2: Manipulating behavioral flexibility (improving reversal learning speed through serial re-**  
123 **versals using shaded tubes) improves flexibility (rule learning and/or switching) and innova-**  
124 **tiveness in a new context (two distinct multi-access boxes and serial reversals on a touchscreen).**

- 125 • **P2:** Individuals that have improved their flexibility on a serial reversal learning task using shaded  
126 tubes (requiring fewer trials to reverse a preference as the number of reversals increases) are faster to  
127 switch between new methods of solving (latency to solve or attempt to solve a new way of accessing  
128 the food [locus]), and learn more new loci (higher total number of solved loci) on multi-access box  
129 flexibility tasks, and are faster to reverse preferences in a serial reversal task using a touchscreen than  
130 individuals in the control group where flexibility has not been manipulated. The positive correlation  
131 between reversal learning performance using shaded tubes and a touchscreen (faster birds have fewer  
132 trials) and the multi-access boxes (faster birds have lower latencies) indicates that all three tests  
133 measure the same ability even though the multi-access boxes require inventing new rules to solve new  
134 loci (while potentially learning a rule about switching: “when an option becomes non-functional, try  
135 a different option”) while reversal learning requires switching between two rules (“choose light gray”  
136 or “choose dark gray”) or learning the rule to “switch when the previously rewarded option no longer  
137 contains a reward”. Serial reversals eliminate the confounds of exploration, inhibition, and persistence  
138 in explaining reversal learning speed because, after multiple reversals, what is being measured is the  
139 ability to learn one or more rules. If the manipulation works, this indicates that flexibility can be  
140 influenced by previous experience and might indicate that any individual has the potential to move  
141 into new environments (see relevant hypotheses in preregistrations on [genetics](#) (R1) and [expansion](#)  
142 (H1)).
- 143 • **P2 alternative 1:** If the manipulation does not work in that those individuals in the experimental  
144 condition do not decrease their reversal speeds more than control individuals, then this experiment  
145 elucidates whether general individual variation in flexibility relates to flexibility in new contexts (two  
146 distinct multi-access boxes and serial reversals on a touchscreen) as well as innovativeness (multi-access  
147 boxes). The prediction is the same as in P2, but in this case variation in flexibility is constrained by  
148 traits inherent to the individual (some of which are tested in McCune KB et al., 2019), which suggests  
149 that certain individuals will be more likely to move into new environments.
- 150 • **P2 alternative 2:** If there is no correlation between reversal learning speed (shaded tubes) and the  
151 latency to solve/attempt a new locus on the multi-access boxes, this could be because the latency  
152 to solve not only measures flexibility but also innovativeness. In this case, an additional analysis is  
153 run with the latency to solve as the response variable, to determine whether the fit of the model (as  
154 determined by the lower AIC value) with reversal learning as an explanatory variable is improved if

155 motor diversity (the number of different motor actions used when attempting to solve the multi-access  
156 box) is included as an explanatory variable (see Diquelou et al., 2015; Griffin et al., 2016). If the  
157 inclusion of motor diversity improves the model fit, then this indicates that the latency to solve a  
158 new locus on the multi-access box is influenced by flexibility (reversal learning speed) and innovation  
159 (motor diversity).

- 160 • **P2 alternative 3:** If there is a negative correlation or no correlation between reversal learning speed  
161 on shaded tubes and reversal learning speed on the touchscreen, then this indicates that it may be  
162 difficult for individuals to perceive and/or understand images on the touchscreen in contrast with  
163 physical objects (shaded tubes) (e.g., O’Hara et al., 2015).

### 164 **H3: Behavioral flexibility within a context is repeatable within individuals.**

165 This hypothesis from the original preregistration is now being treated in a separate manuscript (K. McCune  
166 et al., 2022).

### 167 **H4: Individuals should converge on an epsilon-first learning strategy (learn the correct choice 168 after one trial) as they progress through serial reversals.**

- 169 • **P4:** Individuals prefer a mixture of learning strategies in the first serial reversals (an *epsilon-decreasing*  
170 strategy where individuals explore both options extensively before learning to prefer the rewarded op-  
171 tion, and an *epsilon-first* strategy where the correct choice is consistently made after the first trial),  
172 and then move toward the epsilon-first learning strategy. The epsilon-first strategy works better later  
173 in the serial reversals where the reward is all or nothing because individuals have learned the environ-  
174 ment is changing in predictable ways (Bergstrom & Lachmann, 2004): only one option is consistently  
175 rewarded, and if the reward isn’t in the previously rewarded option, it must be in the other option.
- 176 • **P4 alternative 1:** Individuals continue to prefer a mixture of learning strategies, and/or they do not  
177 converge on the more functional epsilon-first learning strategy, regardless of how many reversals they  
178 participate in. This pattern could suggest that the grackles do not attend to functional meta-strategies,  
179 that is, they do not learn the overarching rule (once food is found in the non-preferred tube, one must  
180 switch to preferring that tube shade), but rather they learn each preference change as if it was new.

## 181 **METHODS**

182 This study is based on a preregistration that received in principle acceptance at PCI Ecology ([PDF](#) version),  
183 which included a description of the analyses we initially planned to perform. In the following, we first outline  
184 the rationale for any changes from the preregistered methods before describing the methods that were used  
185 to derive the results presented here.

### 186 **Changes after pilot data were collected and before the actual data collection began**

- 187 1) We initially (in 2017) set the serial reversal passing criterion as the following. During the data collection  
188 period, the number of trials required to reverse a preference will be documented per bird, and reversals  
189 will continue until the first batch of birds tested reaches an asymptote (i.e., there are negligible further  
190 decreases in the number of trials required to reverse a preference). The number of reversals to reach the  
191 asymptote will be the number of reversals that subsequent birds experience. Due to delays in setting  
192 up the field site, we were only able to test two grackles in early 2018 (January through April) and,  
193 due to randomization, only one (Fajita) was in the experimental condition that involved undergoing  
194 the flexibility manipulation (Empanada was in the control condition). While Fajita’s reversal speeds  
195 generally improved with increasing serial reversals, she never reached an asymptote (which we defined  
196 as passing three consecutive reversals in the same number of trials), even after 38 reversals. These 38  
197 reversals took 2.5 months, which is an impractical amount of time if birds are to participate in the rest  
198 of the test battery (multi-access box, detour, causal cognition, go no-go, reversal on a touchscreen)

199  
200  
201  
202  
203  
204  
205  
206

after undergoing the reversal manipulation (we were initially permitted to keep them in aviaries for up to three months per bird, which we extended to 6 months per bird in Dec 2018). Because our objective in this experiment was to manipulate an individual's flexibility, we decided to revise our serial reversal passing criterion to something more species relevant based on Fajita's serial reversal performance and the performance of seven grackles in Santa Barbara who underwent only one reversal in 2014 and 2015 (Logan, 2016). The **revised serial reversal passing criterion was: passing two reversals in a row at or under 50 trials**. 50 trials is fewer trials than any of the nine grackles required to pass their first reversal (range 70-130), therefore it should reflect an improvement in flexibility.

207 **Changes at the beginning of data collection**

208  
209  
210  
211  
212  
213  
214  
215  
216  
217  
218  
219  
220

2) Reversal learning shaded tube choice criterion. At the beginning of the second bird's initial discrimination in the reversal learning shaded tube experiment (October 2018), we revised the criterion for what counts as a choice from A) the bird's head needs to pass an invisible line on the table that ran perpendicular to the the tube opening to B) the **bird needs to bend its body or head down to look in the tube** (see B demonstrated in Figure 3). Criterion A resulted in birds making more choices than the number of learning opportunities they were exposed to (because they could not see whether there was food in the tube unless they bent their head down to look in the tube) and appeared to result in slower learning. It is important that one choice equals one learning opportunity, therefore we revised the choice criterion to the latter. Anecdotally, this choice matters because the first three birds in the experiment (Tomatillo, Chalupa, and Queso) learned faster than the pilot birds (Empanada and Fajita) in their initial discriminations and first reversals. Thus, it was an important change to make at the beginning of the experiment (after testing the two pilot birds and before collecting any data that were included in analyses).



221

222 **Figure 3.** Tzanatl preciosa bending down to look into the dark gray tube.

223 3) Criterion to pass the control condition: Before collecting experimental data, we set the number of trials  
224 experienced by the birds in the control group as 1100 because this is how many trials it would have  
225 taken the pilot bird in the manipulated group, Fajita, to pass serial reversals 2-17 according to our  
226 revised serial reversal passing criterion. However, after 25 and 17 days (after Tomatillo and Queso's  
227 first reversals, respectively) of testing the first two individuals in the control group, it became apparent  
228 that 1100 trials is impractical given the time constraints for how long we were permitted to keep each  
229 bird temporarily in captivity and would prevent birds from completing the test battery before their  
230 release. Additionally, after revising the choice criterion, it was going to be likely that birds in the  
231 manipulated group would require fewer than 1100 trials to meet the serial reversal passing criterion.  
232 Therefore, reducing the number of trials the control birds experience would result in a better match of  
233 experience with birds in the manipulated group. On 2 November 2018 we **set the number of trials**  
234 **control birds experience after their first (and only) reversal** to the number of trials it requires  
235 the first bird in the manipulated group to pass (the first bird had not passed yet, therefore we did  
236 not yet know what this number was). After more individuals in the manipulated group passed, we  
237 updated this number to the average number of trials to pass. This applied to all birds in the control  
238 condition, except Mofongo. Mofongo (control condition) was a slow participator and would not have  
239 finished his test battery by the time it got too hot to keep birds in the aviaries if we used the current  
240 average number of trials (420). Instead, we matched him with the fastest bird in the manipulated  
241 group (Habanero=290 trials) to make it more likely that Mofongo could get through the rest of the  
242 test battery in time.

#### 243 **Changes in the middle of data collection**

244 4) 10 April 2019, we **discontinued the reversal learning experiment on the touchscreen** because  
245 it appeared to measure something other than what we intended to test and it required a huge time  
246 investment for each bird (which consequently reduced the number of other tests they were available  
247 to participate in). This is not necessarily surprising because this was the first time touchscreen tests  
248 have been conducted in this species, and also the first time (to our knowledge) this particular reversal  
249 experiment has been conducted on a touchscreen with birds. We based this decision on data from  
250 four grackles (2 in the flexibility manipulation group and 2 in the flexibility control group; 3 males  
251 and 1 female). All four of these individuals showed highly inconsistent learning curves and required  
252 hundreds more trials to form each preference when compared to the performance of these individuals  
253 on the shaded tube reversal experiment. It appeared that there was a confounding variable with  
254 the touchscreen such that they were extremely slow to learn a preference as indicated by passing our  
255 criterion of 17 correct trials out of the most recent 20. We did not include the data from this experiment  
256 when conducting the cross-test comparisons in the Analysis Plan section of the preregistration. Instead,  
257 in Supplementary Material 4, we provided summary results for this experiment and, in the Discussion,  
258 qualitatively compared it with performance on the shaded tube reversal test to explain what might  
259 have confounded the touchscreen experiment.

260 5) 16 April 2019, because we discontinued the touchscreen reversal learning experiment, we **added an**  
261 **additional but distinct multi-access box** task, which allowed us to continue to measure flexibility  
262 across three different experiments. There are two main differences between the first multi-access box,  
263 which is made of plastic, and the new multi-access box, which is made of wood. First, the wooden  
264 multi-access box is a natural log in which we carved out 4 compartments. As a result, the apparatus and  
265 solving options are more comparable to what grackles experience in the wild, though each compartment  
266 is covered by a transparent plastic door that requires different behaviors to open. Furthermore, there  
267 is only one food item available in the plastic multi-access box and the bird could use any of 4 loci  
268 to reach it. In contrast, the wooden multi-access box has a piece of food in each of the 4 separate  
269 compartments.

#### 270 **Updates and changes post data collection, pre-data analysis**

271 6) We completed our simulation to explore the lower boundary of a minimum sample size and determined  
272 that **our sample size for the Arizona study site is above the minimum** (see details and code  
273 in Supplementary Material 1; 17 April 2020).

274 7) Please see our Alternative Analyses section in the preregistration where we stated that we would  
275 learn and implement Bayesian models, which resulted in our **changing the analysis for P2** and  
276 that we are replacing this analysis with the new models in the Ability to detect actual effects section  
277 (Supplementary Material 1; 14 May 2020). We also describe in SM1 that we realized that Condition  
278 (manipulated or control) does not need to be a variable in our models because our analyses in P1  
279 demonstrate that the manipulation causally changed reversal speeds, which is the key assumption in  
280 P2.

281 8) We originally planned on testing only **adults** to have a better understanding of what the species is  
282 capable of, assuming the abilities we are testing are at their optimal levels in adulthood, and so we  
283 could increase our statistical power by eliminating the need to include age as an independent variable  
284 in the models. Because the grackles in Arizona were extremely difficult to catch, we ended up testing  
285 two juveniles: Taco and Chilaquile. We did not conduct the full test battery with Taco or put him in  
286 the flexibility manipulation or control groups (he received 1 reversal and then moved on to the next  
287 test) because he was the first juvenile and we wanted to see whether his performance was different  
288 from adult performances. His performances were similar to the adults, therefore we decided to put  
289 Chilaquile in the full test battery. Chilaquile's performances were also similar to the adults, therefore  
290 we decided not to add age as an independent variable in the models to avoid reducing our statistical  
291 power.

292 9) We **removed experimenter as a random effect** from all analyses because the interobserver reli-  
293 ability scores were so high, indicating there was no difference between experimenters, therefore we could  
294 keep our models simpler by leaving this variable out.

295 10) P2 alternative 2: We **used the average latency rather than the number of trials to attempt a**  
296 **new locus** because this would make the model comparable with the model in P2. Using the number  
297 of trials was an artifact from a previous version and we had missed updating this. We omitted the  
298 number of trials to solve a new locus as described in the deviation from the plan in P2 above. We used  
299 a GLM rather than a GLMM because there was only one data point per bird (note that there would  
300 have been only one data point per bird in the preregistration as well, but we didn't realize this until  
301 after in principle acceptance).

302 11) P4 (Aug 2021): The grackles were tested in **10-trial blocks** and not 20-trial blocks as in Federspiel et  
303 al. (2017), which would mean that if there were <20 trials in the last block of a reversal, they would be  
304 omitted from the analysis. Therefore, we changed the block size to 10 trials and adjusted the sampling  
305 blocks to 2-9 correct choices, and the acquisition blocks to 9-10 correct choices using significance levels  
306 in the binomial test as did Federspiel et al. (2017).

### 307 **Changes post data collection, mid-data analysis**

308 12) P2 (April 2020): We realized that the average latency to solve a new locus after solving a different  
309 locus is confounded with the total number of loci solved because the measure of innovation is included  
310 in the definition. Therefore, we removed average latency to solve a locus from analyses so that we  
311 are only examining pure measures of flexibility (average latency to **attempt** to solve) and innovation  
312 (total number of loci solved).

313 13) P2: Removed aviary batch (random variable) from the original model for P2 (Table SM3: Model 1).  
314 Batch ended up confounding the analysis because control and manipulated individuals, while randomly  
315 assigned to these conditions, ended up in particular batches as a result of their willingness to participate  
316 in tests offered during their time in the aviary (Table SM3: Model 3). Several grackles never passed  
317 habituation or training such that their first experiment could begin, therefore we replaced these grackles  
318 in the aviaries with others who were willing to participate. This means that batch did not indicate a



319 particular temporal period. Therefore, we **removed batch from the models** (post data collection,  
320 mid-data analysis).

321 14) P2: When making the bespoke Bayesian models, we realized that we had previously misinterpreted  
322 which variable should be the response variable in this analysis. We originally set the number of trials  
323 to reverse as the response variable, however we should have instead set the number of loci solved as  
324 the response variable and then planned to conduct a second model with the latency to attempt a new  
325 locus as the response variable and number of trials as the explanatory variable. This is because a)  
326 we manipulated the number of trials to reverse, therefore it must be the explanatory variable (Hernán  
327 & Robins, 2006); and b) they should be split into two models, **one each for average latency and**  
328 **number of loci solved**, because of a and because these are two very different relationships that  
329 should be considered in their own models. We also realized that Condition (manipulated or control)  
330 does not need to be a variable in any of our models because our analyses in P1 demonstrate that the  
331 manipulation causally changed reversal speeds, which is the key assumption in P2.

### 332 **Changes post data collection, post-data analysis**

333 15) We present the results from different hypotheses in separate articles: this one, K. McCune et al. (2022),  
334 and Lukas et al. (2022).

### 335 **Sample**

336 Grackles were caught in the wild in Tempe, Arizona, USA for individual identification (colored leg bands  
337 in unique combinations). Some individuals (34: 13 in the control group (they receive 1 reversal; only 11  
338 completed the experiment) and 10 in the flexibility manipulation (they receive multiple reversals; only 8  
339 completed the experiment), and 11 who did not participate enough to enter the experiments) were brought  
340 temporarily into aviaries for testing, and then released back to the wild.

### 341 **Data collection stopping rule**

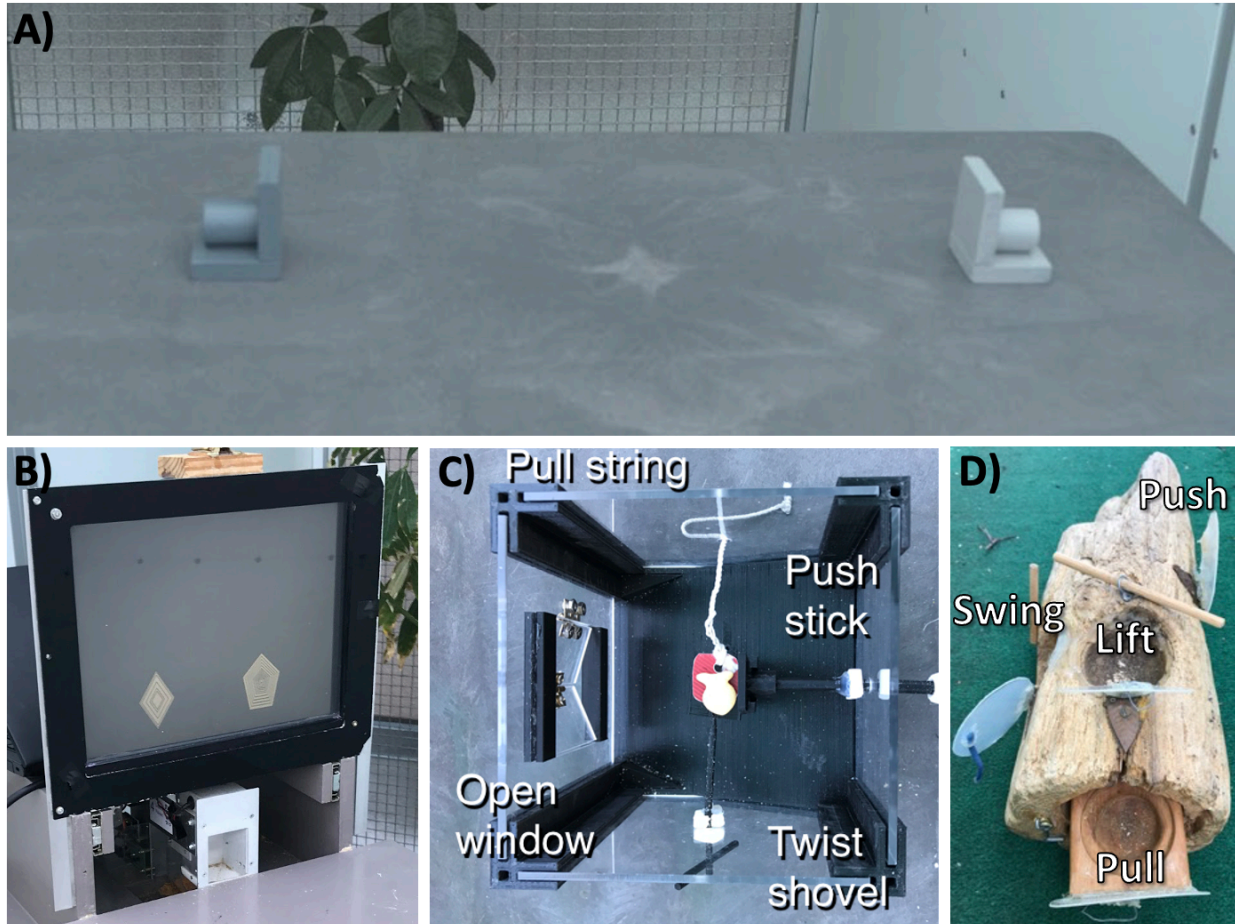
342 We stopped testing birds after we completed two full aviary seasons because the sample size was above  
343 the minimum suggested boundary of 15 (to detect a medium effect size) based on model simulations (see  
344 Supplementary Material 1).

### 345 **Summary of testing protocols (Figure 2)**

- 346 • **Reversal learning with shaded tubes:** One light gray and one dark gray tube were placed such that  
347 the openings were not visible (shades were pseudorandomized for side). One shade always contained a  
348 food reward. The individual had the opportunity to choose to look inside one tube per trial. Once the  
349 individual chose correctly on 17 out of the most recent 20 trials, they were considered to have a shade  
350 preference, and then the food was always placed in the previously non-rewarded shade and the same  
351 passing criterion was used to determine their reversal learning performance. Individuals were randomly  
352 placed in the manipulated condition (serial reversals until they passed two consecutive reversals in 50  
353 trials or less) or the control condition (receive only one reversal and then a similar number of total  
354 trials to the manipulated individuals, but with two yellow tubes, both of which always had food).
- 355 • **Plastic multi-access box:** This was a puzzlebox made of plexiglas and plastic, which contained one  
356 piece of food on a post in the center of the box. The box was placed in the aviary for up to 15 minutes  
357 per trial. Each plexiglas wall had one option (locus) for retrieving the food, but each option required  
358 a different method for obtaining the food. The individual had the opportunity to attempt (touch, but  
359 not obtain the food) or solve a locus. Once a locus was used successfully three times to get the food, it  
360 was considered solved and rendered non-functional in subsequent trials. The experiment ended when  
361 an individual solved all four loci or if they did not interact with or successfully solve a locus in three  
362 consecutive trials.

363  
364  
365  
366  
367  
368  
369  
370  
371  
372

- **Wooden multi-access box:** This was a puzzlebox carved from a log to have four loci containing a food item. Each locus required a different motor action to solve. Three loci were covered with a plastic door on a hinge and one locus was a drawer that must be pulled out. Trials lasted for up to 15 minutes. The passing criterion and experiment ending criteria were the same as for the plastic multi-access box.
- **Reversal learning of shapes on a touchscreen:** This is the same experimental design as with the shaded tubes, except it was carried out on a touchscreen computer where the individual was presented with two white symbols that differed in shape (pentagon or diamond). Touching the screen over the rewarded shape resulted in food dropping from a food hopper into a dish accessible to the grackle, while touching the screen over the non-rewarded shape resulted in no food and a longer inter-trial interval.



373

374 **Figure 2.** The experimental apparatuses: reversal learning using dark gray and light gray tubes or two  
375 different shapes on a touchscreen, and the wooden and plastic multi-access boxes (MAB). The wooden MAB  
376 has four loci, each containing food and each locus has a distinct way of being opened: lift up flap (A), swing  
377 open flap (B), pull out drawer (C), or push in flap (D). The plastic MAB has four loci that all provide access  
378 to one piece of food and each locus has a distinct way of being opened: open the window (left side), pull the  
379 string (top side), push the shovel (right side), or twist the shovel (bottom side).

### 380 Open materials

381  
382  
383

- [Design files](#) for the plastic multi-access box: 3D printer files and laser cutter files
- [Testing protocols](#) for all experiments: shaded tube reversal learning, plastic multi-access box, wooden multi-access box, and touchscreen reversal learning

384 **Open data**

385 Data are publicly [available](#) at the Knowledge Network for Biocomplexity (C. Logan et al., 2023).

386 **Randomization and counterbalancing**

387 H1: Subjects were randomly assigned to the manipulated or control group. In the reversal learning trials,  
388 the rewarded option is pseudorandomized for side (and the option on the left is always placed first). Pseudo-  
389 randomization consisted of alternating location for the first two trials of a session and then keeping the same  
390 shade on the same side for at most two consecutive trials thereafter. A list of all 88 unique trial sequences for  
391 a 10-trial session, following the pseudorandomization rules, was generated in advance for experimenters to  
392 use during testing (e.g., a randomized trial sequence might look like: LRLRLRLRLR, where L and R refer to  
393 the location, left or right, of the rewarded tube). Randomized trial sequences were assigned randomly to any  
394 given 10-trial session using a random number generator (random.org) to generate a number from 1-88. The  
395 only exception to this randomization was when an individual exhibited a side bias (choosing one side 4 or  
396 more trials in a row). In these cases, we stopped the current random numbers for side and started putting the  
397 rewarded shade on the non-preferred side as much as possible while still following the pseudorandomization  
398 rules until the individual stopped exhibiting a side bias.

399 **ANALYSES**

400 Analyses were conducted in R (current version 4.1.2, R Core Team, 2017), using several R packages: kable-  
401 Extra (Zhu, 2021), MCMCglmm (Hadfield, 2010), MuMIn (Bartoń, 2020), rethinking (McElreath, 2020),  
402 stan (Stan Development Team, 2020), formatR (Xie, 2019), Rstudioapi (Ushey et al., 2020), rcpp (Eddel-  
403 buettel & François, 2011), ggplot2 (Wickham, 2016), knitr (Xie, 2013, 2017, 2018), dplyr (Wickham et al.,  
404 2021), cmdstanr (Gabry & Češnovar, 2021), cowplot (Wilke, 2017), reactable (Lin, 2020), DHARMa (Hartig,  
405 2019), and lme4 (Bates et al., 2012; Bates et al., 2015).

406 **Unregistered analyses:** We conducted unregistered interobserver reliability analyses on the video and live  
407 coding of the response variables. Scores indicated that the response variables are repeatable to a high or  
408 extremely high degree given our instructions and training for coders (see Supplementary Material 2).

409 **Data checking**

410 The data were checked for overdispersion, underdispersion, zero-inflation, and heteroscedasticity with the  
411 DHARMa R package (Hartig, 2019).

412 **P1: Negative relationship between the number of trials to reverse a preference and the number**  
413 **of reversals?**

414 **Analysis:** Response variable: Number of trials to reverse a preference. We use a sliding window to look  
415 at the most recent 10 trials for a bird, regardless of when the testing sessions occurred. Explanatory  
416 variable: reversal number. Random variables: batch (batch is a test cohort, consisting of 8 birds being  
417 tested simultaneously and there were multiple batches included in the analysis) and ID (random effect  
418 because there were repeated measures on the same individuals). A Generalized Linear Mixed Model (GLMM,  
419 MCMCglmm function, MCMCglmm package, Hadfield, 2010) was used with a Poisson distribution and log  
420 link using 300,000 iterations with a thinning interval of 500, a burnin of 90,000, and minimal priors ( $V=1$ ,  
421  $\nu=0$ ) (Hadfield, 2014). We ensured the GLMM showed acceptable convergence (lag time autocorrelation  
422 values  $<0.01$ , Hadfield, 2010), and adjusted parameters as necessary.

423 We did not need a power analysis to estimate our ability to detect actual effects because, by definition, the  
424 individuals that complete this experiment must get faster at reversing in order to pass the stopping criterion  
425 (two consecutive reversals in 50 trials or less). According to previous grackle data (from the pilot birds, and  
426 from Santa Barbara Logan, 2016), the fastest grackle passed their first reversal in 70 trials, which means  
427 that passing our serial reversal stopping criterion would require them to have improved their passing speed.

428 **Unregistered analyses:** We evaluated whether the individuals in both conditions (manipulated and con-  
429 trol) required a similar number of trials to pass their first reversal (dependent variable: trials to reverse in  
430 first reversal, explanatory variable: condition, random variables: ID and batch; Table 1), and their last rever-  
431 sal (dependent variable: trials to reverse in last reversal, explanatory variable: condition, random variables:  
432 ID and batch; Table 3).

## 433 **P2: Serial reversal improves rule switching and innovativeness**

434 **Analyses:** One model was run per response variable: average latency to attempt to solve a new locus after  
435 solving a different locus, and total number of loci solved. Explanatory variable: Number of trials to reverse  
436 a preference in the last reversal.

437 The model for the number of loci solved takes the form of:

$$438 \text{locisolved}_{i,j} \sim \text{Binomial}(4, p) \text{ [likelihood]},$$

$$439 \text{logit}(p) \sim \alpha + \beta \text{trials}_{i,j} \text{ [model]},$$

440 where  $\text{locisolved}_{i,j}$  is the number of loci solved on the multi-access box, 4 is the total number of loci on the  
441 multi-access box,  $p$  is the probability of solving any one locus across the whole experiment,  $\alpha$  is the intercept,  
442  $\beta$  is the expected amount of change in  $\text{locisolved}_{i,j}$  for every one unit change in  $\text{trials}_{i,j}$ , and  $\text{trials}_{i,j}$  is the  
443 number of trials to reverse a shade preference. See Supplementary Material 1 for more model details.

444 The model for the latency to switch options takes the form of:

$$445 \text{latency}_{i,j} \sim \text{gamma-Poisson}(\lambda_{i,j}, \phi) \text{ [likelihood]},$$

$$446 \log(\lambda_{i,j}) \sim \alpha + \beta \text{trials}_{i,j} \text{ [model]},$$

447 where  $\text{latency}_{i,j}$  is the average latency to attempt a new locus on the multi-access box,  $\lambda_i$  is the rate  
448 (probability of attempting a locus in each second) per bird (and we take the log of it to make sure it is  
449 always positive; birds with a higher rate have a smaller latency),  $\phi$  is the dispersion of the rates across birds,  
450  $\alpha$  is the intercept for the rate,  $\beta$  is the expected amount of change in the rate of attempting to solve in  
451 any given second for every one unit change in trials, and trials is the number of trials to reverse a shade  
452 preference. Note that a gamma-Poisson distribution is also known as negative binomial. See Supplementary  
453 Material 1 for more model details.

454 Note: As originally planned, we replaced the GLMs and GLMMs in May 2020 with more powerful models  
455 after learning how to make bespoke Bayesian models from McElreath (2016). We made these models before  
456 analyzing the actual data (14 May 2020).

457 **Unregistered analysis:** Because the wooden multi-access box was added after in principle recommendation,  
458 we conducted an unregistered analysis to determine whether the plastic and wooden multi-access box results  
459 correlated with each other, which would indicate that these tests are interchangeable. We found that they  
460 did not statistically significantly correlate with each other on either variable measured: the average latency  
461 to attempt a new locus (switching; Pearson’s  $r=0.74$ , 89% confidence level=0.02-0.95,  $t=2.18$ ,  $df=4$ ,  $p=0.09$ ,  
462  $n=6$ ) or the total number of loci solved (problem solving; Pearson’s  $r=0.51$ , 89% confidence level=0.03-0.80,  
463  $t=1.86$ ,  $df=10$ ,  $p=0.09$ ,  $n=12$ ). Therefore, while the performance on the two multi-access boxes might not  
464 be completely independent as indicated by the high  $r$  values, the two boxes appear not to be completely  
465 interchangeable either as indicated by the lack of statistical significance and high uncertainty in the  $r$  values.  
466 We therefore analyzed the plastic and wooden multi-access boxes separately.

467 Post-data collection, we added an additional unregistered analysis comparing first versus last reversal perfor-  
468 mance for the individuals in the manipulated group (see r code chunk “posthoc\_conditionalimprovement”  
469 at the rmd for model details).

## 470 **P2 alternative 2: Additional analysis: latency and motor diversity**

471 **Analyses:** We ran one model per response variable: average latency to attempt a new locus on the multi-  
472 access boxes, and number of trials to solve (meet criterion) a new locus on the multi-access boxes. Explan-  
473 atory variables: Number of trials to reverse a preference in the last reversal that an individual participated

474 in, the number of different motor actions used when attempting to solve the multi-access boxes (motor  
 475 diversity). A General Linear Model (GLM; glm function) was used with a Poisson distribution and log link.

476 **P4: Learning strategies (for birds in the manipulated group only)**

477 **Analysis 1 (qualitative):** Learning strategies were identified by matching them to the two known approx-  
 478 imate strategies of the contextual, binary multi-armed bandit: epsilon-first and epsilon-decreasing (McIn-  
 479 ernerney, 2010; as in Logan, 2016). We used the criterion for the epsilon-first strategy of learning the correct  
 480 choice after one trial and then choosing correctly thereafter. Other patterns were classified as the epsilon-  
 481 decreasing strategy where individuals gradually increase their number of successes as the number of trials  
 482 increases. This method of qualitative inspection of learning curves is standard for this type of learning strat-  
 483 egy assessment (McInerney, 2010). The variable for visual inspection was the proportion of correct choices  
 484 in a non-overlapping sliding window of 4-trial bins across the total number of trials required to reach the  
 485 criterion of 17/20 correct choices per individual.

486 **Analysis 2 (quantitative):** We then quantitatively determined to what degree each bird used the explo-  
 487 ration versus exploitation strategy using methods in Federspiel et al. (2017) by calculating the number of  
 488 10-trial blocks where birds were choosing “randomly” (2-9 correct choices; called sampling blocks; akin to  
 489 the exploration phase above) and dividing it by the total number of blocks to reach criterion per bird. This  
 490 ratio was also calculated for “acquisition” blocks where birds made primarily correct choices (9-10 correct  
 491 choices; akin to the exploitation phase above). These ratios, calculated for each bird for their serial reversals,  
 492 quantitatively discern the exploration from the exploitation phases.

493 **RESULTS**

494 Although 22 grackles completed their initial shaded tube discrimination, only 20 grackles participated in  
 495 one or more reversal (Table SM5). The rest of the tests began only after a bird’s reversal experiment was  
 496 complete (C. Logan et al., 2023).

497 **P1: Reversal speed gets faster with serial reversals**

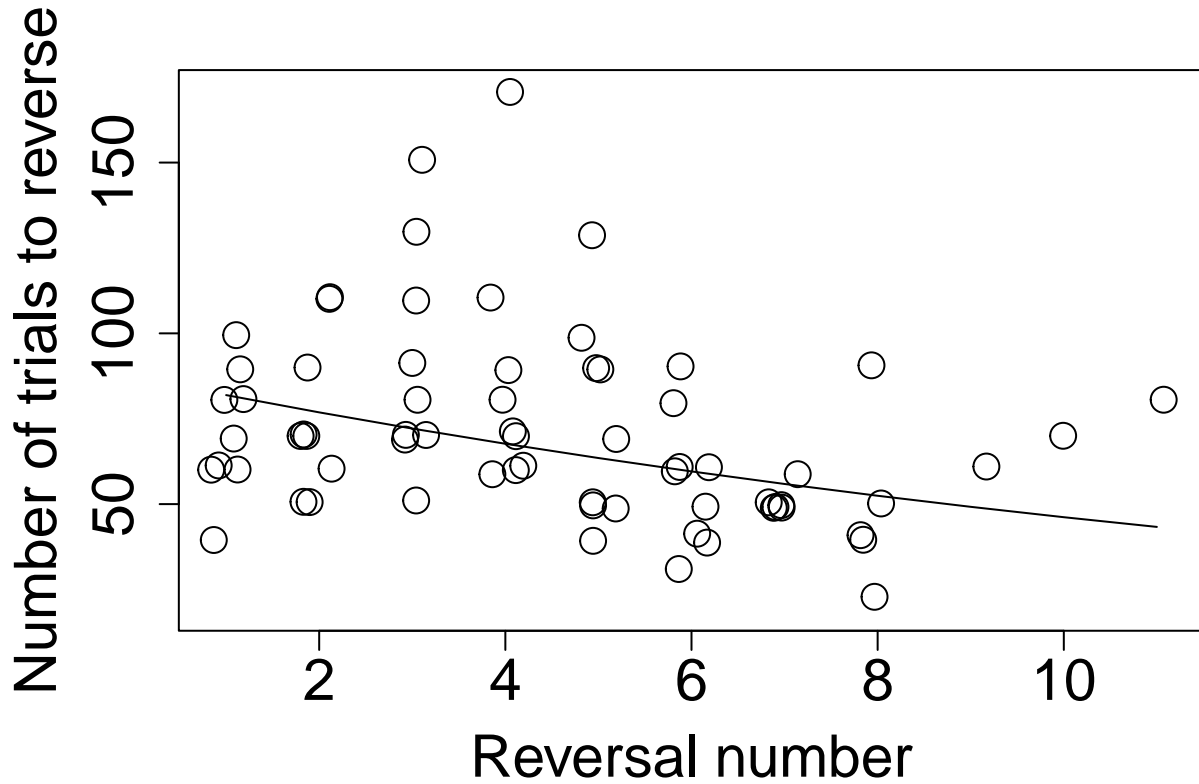
498 The birds in the manipulated group required a similar number of trials during their first reversal (R1 me-  
 499 dian=75 trials) as the birds in the control group needed during their first and only reversal (R1 median=70  
 500 trials) (see unregistered analysis in Table 1). The manipulated birds improved during the reversal manip-  
 501 ulation to a median of 40 trials in their last reversal: there was a significant negative correlation between  
 502 the number of trials to reverse (average=71 trials, standard deviation (sd)=28, Table 2) and the reversal  
 503 number for those grackles in the flexibility manipulation condition (n=9, which included Memela who did  
 504 not pass the manipulation condition of passing two consecutive reversals in 50 trials or less; Figure 4).

505 **Table 1.** Unregistered analysis: the number of trials to reverse in the first reversal is similar between the  
 506 manipulated and control groups.

	Posterior mean	Lower 89 percentile compatibility interval (5.5%)	Upper 89 percentile compatibility interval (94.5%)	Effective sample size	pMCMC	Significance code: **=0.01
Intercept	4.29	4.12	4.46	420	<0.002	**
Manipulation Condition	-0.08	-0.27	0.11	420	0.46	

509 **Table 2.** In the manipulated birds, the number of trials to reverse decreases with increasing reversal number.

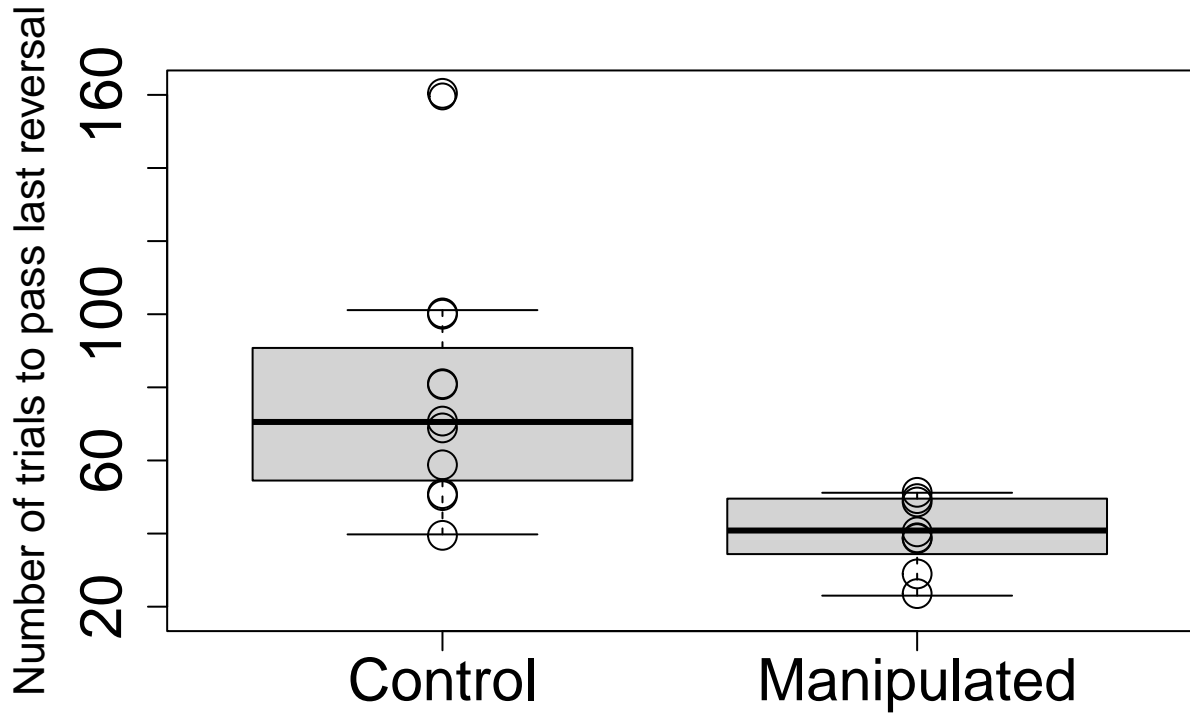
	Posterior mean	Lower 89 percentile compatibility interval (5.5%)	Upper 89 percentile compatibility interval (94.5%)	Effective sample size	pMCMC	Significance code: **=0.01
Intercept	4.44	4.31	4.62	420	<0.002	**
Reverse Number	-0.06	-0.10	-0.03	420	<0.002	**



512

513 **Figure 4.** Individuals in the manipulated condition (who received serial reversals) linearly decreased their  
514 reversal passing speeds with increasing reversal number (n=9 grackles).

515 **Unregistered analysis 1:** There was additionally a difference between manipulated and control reversal  
516 speeds when comparing their last reversals (Figure 5; for the control birds, their last reversal was their first  
517 reversal; Table 3). This analysis includes 19 grackles (8 manipulated condition - only those who actually  
518 passed the manipulation, 11 control condition) who had an overall average of 62 trials in their last reversal  
519 (sd=32).



520

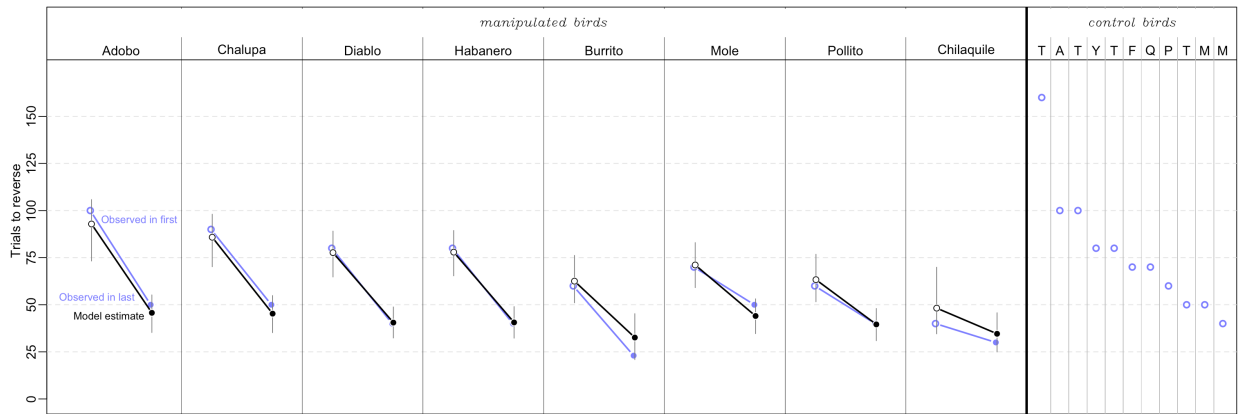
521 **Figure 5.** Individuals in the manipulated condition (who received serial reversals) passed their last reversal  
 522 in fewer trials than individuals in the control condition (who only received 1 reversal). n=19 grackles:  
 523 11=control, 8=manipulated.

524 **Table 3.** Individuals in the manipulated condition pass their last reversal in fewer trials than control  
 525 individuals.

	Posterior mean	Lower 89 percentile compatibility interval (5.5%)	Upper 89 percentile compatibility interval (94.5%)	Effective sample size	pMCMC	Significance code: **=0.01
Intercept	4.28	4.08	4.48	420	<0.002	**
Reverse Number	-0.51	-0.81	-0.22	420	0.010	**

526

528 **Unregistered analysis 2:** A pooled model of performance across all reversals estimates that birds can  
 529 expect to improve by about 30 trials (89% percentile interval (PI): 25-36; Table SM3: Model 15) after  
 530 completing the serial reversals. While all manipulated birds improved, those birds that were already fast to  
 531 reverse in their first reversal improved less than the birds that required many trials to reverse in their first  
 532 reversal (posterior peak indicates a correlation of +0.64, with highest posterior density intervals (HPDI) all  
 533 positive, between the first reversal value and the improvement achieved by the last reversal; Table SM3:  
 534 Model 16). However, the birds who were the fastest in the first reversal, were also the fastest in the last  
 535 reversal, but the difference between the slower and faster reversers is reduced (Figure 6).





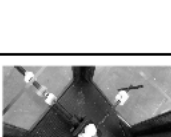

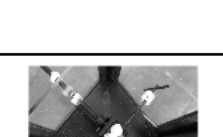
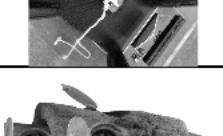
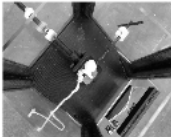

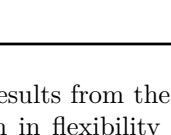

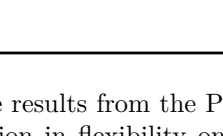
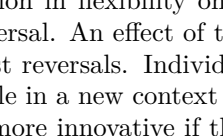
536

537 **Figure 6.** All eight manipulated birds (8 panels on the left) needed fewer trials to reverse in their last  
 538 reversal than in their first. Their improvement depended on their starting value, with steeper slopes for  
 539 those birds that needed more trials to reverse in the first reversal (blue = observed values and changes,  
 540 black = model estimates). However, birds who needed more trials in the first reversal did not completely  
 541 catch up, such that the birds that needed more trials in their first reversal also needed more trials in their  
 542 last reversal relative to other grackles. The panel on the right shows the observed values (which were almost  
 543 exactly the same as the model estimates) for the control birds who received only one reversal. The letters  
 544 in the columns for the control birds are the first letter of their name (from left to right: Taquito, Adobo,  
 545 Tapa, Yuca, Taco, Fideo, Queso, Pizza, Tomatillo, Marisco, Mofongo).

546 **P2: Serial reversals improve rule switching and innovativeness on the MAB**

547 To determine whether the serial reversal manipulation affected flexibility generally, we compared three  
 548 measures of performance (the number of trials to reverse a preference in the first and last shade reversal,  
 549 performance of the manipulated group relative to the control group) to the speed of solution switching on  
 550 two multi-access boxes. Furthermore, we assessed whether flexibility measured through these serial reversals  
 551 related to innovativeness by comparing performance to the number of loci solved on the multi-access boxes.  
 552 The results for each of these comparisons are described in detail below and an overview is provided in Figure  
 553 7.



P2: How does flexibility, measured via performance on serial reversals, relate to flexibility in another context and innovativeness?	Flexibility (serial reversals)		
	First Reversal	Last Reversal	Manipulated relative to Control
Flexibility in a new context (locus switching)	 + 11	 + 9	 + 10
	 - 14	 0 12	 0 13
Innovativeness (locus solving)	 0 5	 +* 2	 0 4
	 0 8	 0 6	 + 7

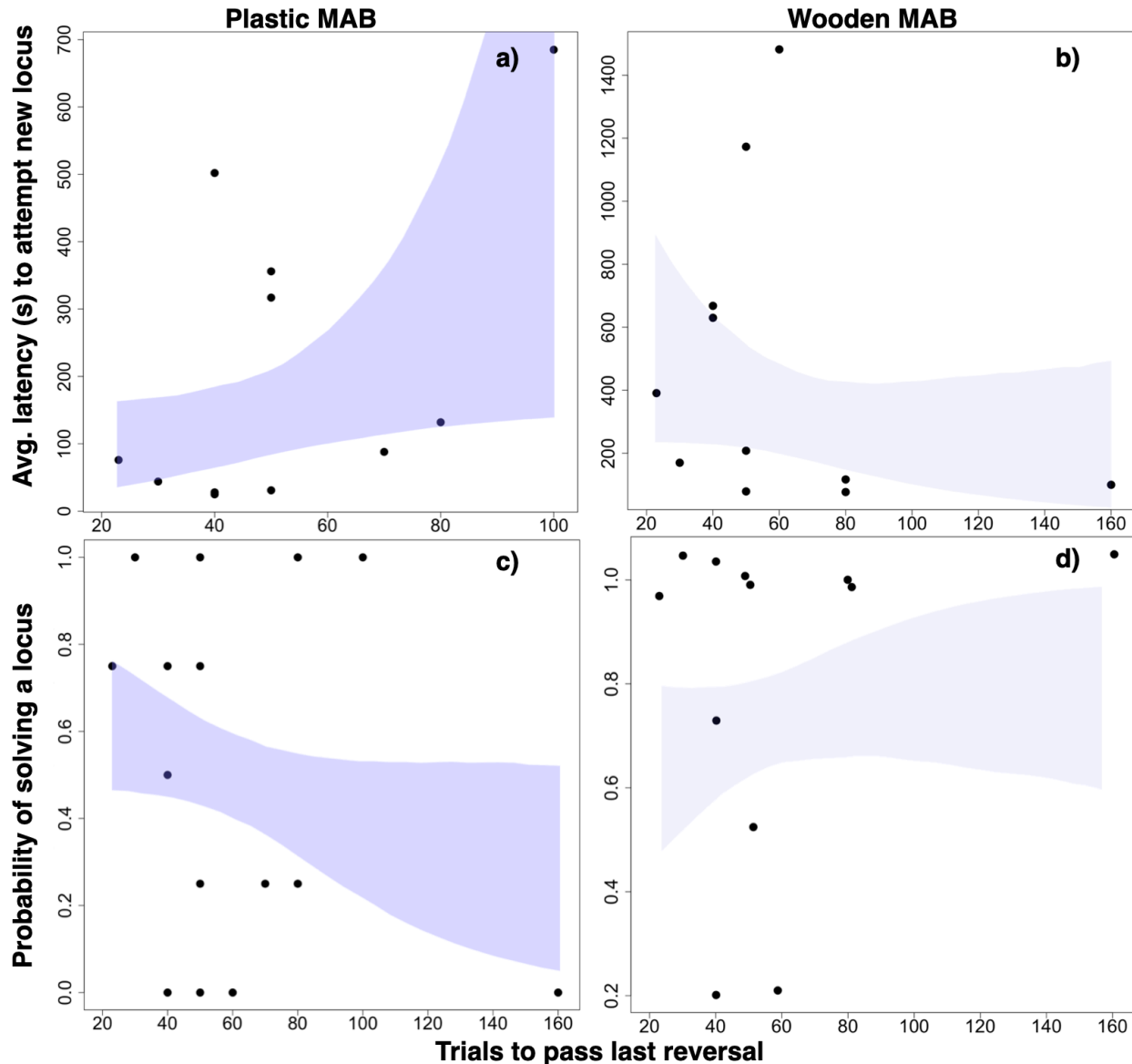
554

555 **Figure 7.** Overview of the results from the P2 analyses with the multi-access boxes (plastic and wooden).  
556 An effect of natural variation in flexibility on performance on the multi-access box tasks would result in  
557 correlations in the first reversal. An effect of the flexibility manipulation would result in a change in corre-  
558 lations from the first to last reversals. Individuals are more flexible if they require fewer trials to pass the  
559 serial reversals, more flexible in a new context if they have shorter latencies to switch to a new locus on the  
560 multi-access box, and are more innovative if they solve more loci on the multi-access box. A plus sign (+)  
561 indicates that the two abilities are positively correlated, a minus sign (-) that they are negatively correlated,  
562 and a 0 indicates no correlation between the two abilities (note that the correlation between the variables  
563 that reflect the abilities for innovativeness have the opposite sign because individuals with more flexibility  
564 need fewer trials in the reversal learning experiment). The asterisk (\*) indicates that a small sample size  
565 decreases the reliability of this result. The number in each cell indicates which model in Table SM3 shows  
566 the model outputs for this result.

567 **Rule switching: Latency to attempt a new locus on the multi-access box (plastic) ~ trials to**  
568 **reverse**

569 Grackles that were faster to reverse a preference in their **last reversal** (average=52 trials, sd=23), where  
570 grackles in the control condition received only one reversal which served as their first and last reversal, were  
571 also faster to attempt to solve a new locus on the plastic multi-access box (after just having passed criterion

572 on a different locus; average=208 seconds, sd=226; Figure 8a; Table SM3: Model 9; n=11 grackles: 6 in  
 573 manipulated condition, 5 in control condition; 6 subjects completed this experiment but solved 0 loci or 1  
 574 locus and so did not have switching times). We also found that individuals in the flexibility manipulation  
 575 had faster switch latencies than those in the control condition (Table SM3: Model 10). Lastly, there was a  
 576 positive correlation between the number of trials to reverse in the **first reversal** (average=70 trials, sd=21)  
 577 and the average switch latency on the plastic multi-access box (Table SM3: Model 11). A correlation was  
 578 determined to be present if the compatibility interval for the slope (b) in the model output did not cross  
 579 zero (Table SM3). This criterion was used throughout the analyses for P2.



580

581 **Figure 8.** The average latency (seconds) to attempt to solve a different locus after having previously  
 582 successfully solved a locus on a) the plastic multi-access box (MAB) is positively correlated with the number  
 583 of trials to pass their last reversal (n = 11 grackles), but on b) the wooden MAB it is not correlated with  
 584 the number of trials to pass their last reversal (n = 11 grackles). Additionally, the probability of solving a  
 585 locus on c) the plastic MAB is negatively correlated with the number of trials to pass their last reversal (n  
 586 = 15 grackles), but on d) the wooden MAB it is not correlated with the number of trials to pass their last  
 587 reversal (n = 12 grackles, estimate of slope includes zero). Shading represents the 89 percentile compatibility  
 588 intervals and darker shading indicates relationships that were found.

589 **Rule switching: Latency to attempt a new locus on the multi-access box (wooden) ~ trials to**  
590 **reverse (unregistered analysis)**

591 There was no correlation between the number of trials to reverse a preference in their **last reversal** (av-  
592 erage=60 trials, sd=38) and the latency to attempt to solve a new locus on the wooden multi-access box  
593 (after just having passed criterion on a different locus; average=463 seconds, sd=481; Figure 8b; Table SM3:  
594 Model 12; n=11 grackles: 5 in manipulated condition, 6 in control condition; Diablo also completed this  
595 experiment and solved 1 locus, but did not attempt another locus after that, thus he does not have any  
596 switching times to analyze). We additionally found that there was no difference in the average latency to  
597 switch between individuals in the flexibility manipulation and those in the control condition (Table SM3:  
598 Model 13). There was a negative correlation between the number of trials to reverse in the **first reversal**  
599 (average=73 trials, sd=34) and the average switch latency on the multi-access box (Table SM3: Model 14).

600 **Innovativeness: Number of loci solved on the multi-access box (plastic) ~ trials to reverse**

601 Grackles that were faster to reverse a preference in their **last reversal** (average=62 trials, sd=34) solved  
602 more loci on the plastic multi-access box (average=2 loci, sd=1.6; Figure 8c; Table SM3: Model 2; n=15  
603 grackles: 6 in manipulated condition, 9 in control condition; this number excludes Mole and Habanero who  
604 were, due to experimenter error, given the fully put together box during habituation and could have learned  
605 how to solve the loci at that time). There was no correlation between the number of loci solved and which  
606 reversal condition a grackle was randomly assigned to (Table SM3: Model 4). There was also no correlation  
607 between the number of trials to reverse in the **first reversal** (average=75 trials, sd=31) and the number of  
608 loci solved on the multi-access box (Table SM3: Model 5).

609 **Innovativeness: Number of loci solved on the multi-access box (wooden) ~ trials to reverse**  
610 **(unregistered analysis)**

611 The compatibility interval for the estimate for the association (mean beta -0.41) between the number of  
612 loci solved on the wooden multi-access box (average=3.2, sd=1.3) and the number of trials to reverse a  
613 preference in their **last reversal** (average=59 trials, sd=38) crossed zero (Figure 8d; Table SM3: Model  
614 6; n=12 grackles: 6 in manipulated condition, 6 in control condition). This could mean that there is no  
615 association, however simulations in Supplementary Material 1 showed that we would not be able to reliably  
616 distinguish whether a small effect is different from zero with our sample size (with a simulated beta of -1 and  
617 a sd in the number of trials >10, the compatibility interval of the estimate crossed zero in all simulations;  
618 Table SM1.2). We did find a correlation between the number of loci solved and which reversal condition a  
619 grackle was randomly assigned to, indicating the reversal manipulation appears to have affected performance  
620 on the wooden multi-access box. The model estimates that manipulated birds solved on average 1.2 more  
621 loci than birds in the control condition (Table SM3: Model 7, wooden; 89% compatibility intervals=0.34-  
622 2.14; n=12 grackles: 6 in manipulated condition, 6 in control condition). However, there is no association  
623 between the number of trials to reverse in the **first reversal** (average=74 trials, sd=34) and the number of  
624 loci solved on the multi-access box (Table SM3: Model 8, wooden).

625 **P2 alternative 2 (additional analysis): Latency and motor diversity**

626 Because there was no correlation between the number of trials to reverse in the last reversal and the latency to  
627 attempt a different locus on the wooden multi-access box, we conducted this additional analysis to determine  
628 whether the model fit was improved when adding the number of motor actions as an explanatory variable.  
629 Adding the number of motor actions (wooden: average=13, sd=4) did not improve the model fit when  
630 examining the relationship between the latency to switch loci on the wooden multi-access box (average=463,  
631 sd=481) and the number of trials to reverse in the last reversal (average=60, sd=38) because the Akaike  
632 weights were similar for both models (n=11 grackles: 5 in the manipulated group, 6 in the control group;  
633 Table 4).

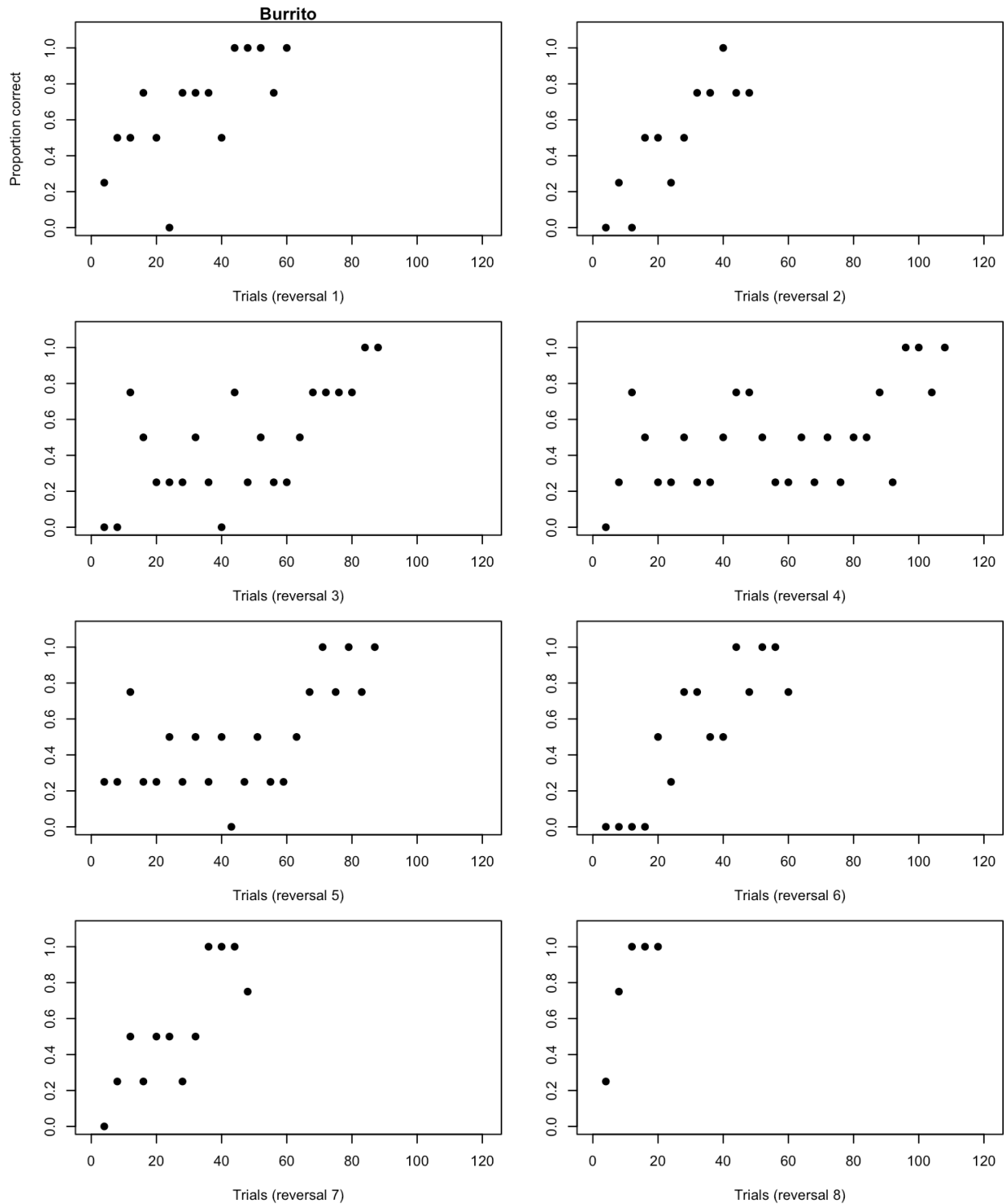
634 **Table 4.** Adding the number of motor actions used to the analysis of the average latency to attempt a  
635 new option on the wooden multi-access box and the number of trials to reverse in the last reversal does not

improve the model fit. Each row represents one model that includes different independent variables (motor actions and/or trials last reversal).

	Intercept	Motor actions (wooden)	Trials last reversal	df	log likelihood	AICc	delta	weight
638	463.2	NA	NA	2	-83.025	171.6	0.00	0.674
	934.6	-35.28	NA	3	-82.477	174.4	2.83	0.164
	665.8	NA	-3.362	3	-82.631	174.7	3.14	0.140
639	1250.0	-40.68	-4.040	4	-81.850	178.4	6.82	0.022

#### 640 **P4: Serial reversal learning strategy**

641 **Analysis 1 (qualitative):** Using the criterion for the epsilon-first strategy of learning the correct choice  
 642 after one trial and then choosing correctly thereafter, no grackle in this study used this strategy in any  
 643 reversal. All grackles used an epsilon-decreasing strategy in all reversals (Figure 9 and Supplementary  
 644 Material 6). We use Burrito's figures to illustrate the epsilon-decreasing strategy (Figure 9): the proportion  
 645 of trials he gets correct wanders up and down (epsilon-decreasing) until an asymptote at 0.8 is reached and  
 646 held.



647

648 **Figure 9.** Burrito's proportion of trials correct by trial number and reversal showing the epsilon-decreasing  
 649 learning strategy where options are explored before forming a preference.

650 **Analysis 2 (quantitative):** We additionally quantitatively determined to what degree each bird used  
 651 the exploration versus exploitation strategy using methods in Federspiel et al. (2017) by calculating the  
 652 number of 10-trial blocks where birds were choosing "randomly" (2-9 correct choices; called sampling blocks;  
 653 akin to the exploration strategy) divided by the total number of blocks to reach criterion per bird. This

654 ratio was also calculated for “acquisition” blocks where birds made primarily correct choices (9-10 correct  
655 choices; akin to the exploitation strategy). There was no correlation between exploration (sampling ratio) or  
656 exploitation (acquisition ratio) and reversal number (sampling: reversal estimate=-0.09, SE=0.11, z=-0.86,  
657 p=0.39; acquisition: reversal estimate=0.00, SE=0.00, z=0, p=1.00), indicating that the grackles did not  
658 use a particular strategy earlier or later in their serial reversals.

## 659 DISCUSSION

660 We conducted a controlled experiment to evaluate whether serial reversal learning affected flexibility and  
661 innovativeness in new contexts. We found that the number of trials to reverse decreased with increasing  
662 reversal number, and, when examining last reversals, there was a difference between the manipulated and  
663 control groups. This indicates that the flexibility manipulation was effective in that it improved reversal  
664 learning speeds, suggesting that these individuals shifted toward a “win-stay, lose-shift” rule to learn to  
665 reverse faster after more experience with reversing (Spence, 1936; J. Warren, 1965; J. M. Warren, 1965).  
666 The manipulated individuals who increased their reversal learning speed, were then apparently able to apply  
667 this to a new context, which resulted in better performance when compared with control individuals who  
668 did not have the opportunity to learn. Previous research has also exploited the fact that most individuals  
669 can learn to learn and have used serial reversals to show that such experience usually improves performance  
670 when transferring to reversals involving different stimuli (e.g., visual vs. spatial, visual vs. visual in a new  
671 combination) (Rayburn-Reeves et al., 2013; Schusterman, 1962; J. Warren, 1965, 1966).

672 While performance differed between the two multi-access boxes, the serial reversal flexibility manipulation  
673 did affect flexibility in a new context, as well as innovativeness (Figure 7). Grackles that were faster to  
674 reverse a preference in their first and last reversals, and those in the manipulated condition, were also faster  
675 to attempt to solve a new locus on the **plastic** multi-access box. Similarly, the flexibility manipulation  
676 affected innovativeness because grackles in the manipulated condition solved on average 1.2 more loci on  
677 the **wooden** multi-access box than those birds in the control condition and there was a negative correlation  
678 between the number of loci solved on the **plastic** multi-access box and the number of trials to reverse in  
679 the last reversal. That our results were not consistent across first reversal, last reversal, and condition  
680 (Figure 7) on the two different multi-access boxes could be due to the small sample sizes because even in  
681 the control group there were several individuals who solved their first and only reversal in very few trials.  
682 Because of the variation in our small sample (Taquito was by far the slowest to reverse a preference), we  
683 conducted a validation check to determine whether removing a bird from the data set changed the model  
684 results. Removing either Taquito or a random bird from the data set changed the conclusions for one of the  
685 three models (Model 2, but not Models 6 or 12). This change in results after removing a data point indicates  
686 that we should be less confident in the conclusion that individuals who are faster to reverse a preference in  
687 their last reversal also solved more loci on the plastic multi-access box. However, it did not matter whether  
688 we removed Taquite, the slowest performer, or a random bird, indicating that this outlier did not drive the  
689 results but rather that the result is constrained by our small sample size. In the cases where there was no  
690 correlation between loci solved and reversal performance, it is possible that the effect size was too small  
691 for us to have the power to detect (Figure 7). Furthermore, the lack of correlation between the number of  
692 trials to reverse in the first reversal and the number of loci solved on either multi-access box indicates that  
693 flexibility is not an inherently utilized tool, but one that is shaped by experience. If it was an inherently  
694 utilized tool, the variation in the number of trials to complete first reversals would likely have resulted in a  
695 correlation with the number of loci solved.

696 Our results are in contrast with previous research on the correlation between flexibility performance on serial  
697 reversals and innovation: Indian mynas that were faster to reverse, were slower to innovate (Griffin et al.,  
698 2013). However, the Griffin et al. (2013) investigation was designed to evaluate the correlation between  
699 the variables and not whether manipulating flexibility using serial reversals influenced innovativeness. This  
700 difference could explain the differing results because correlational research can become noisy if there are  
701 unmeasured variables, which is something that a manipulation can help reduce. Other potential reasons  
702 for the difference in results could include using different experimental designs, and/or different serial re-  
703 versal passing criteria (Griffin et al., 2013 used a preset number of reversals that resulted in a maximum

704 of four reversals), inherent species differences, or needing a larger sample size to help reduce noise in a  
705 non-manipulative experiment.

706 None of the flexibility manipulated individuals converged on using an epsilon-first learning strategy (learn  
707 the correct choice after one trial) as they progressed through serial reversals. All used the epsilon-decreasing  
708 strategy (explore options before forming a preference) throughout their reversals. Additionally, no grackle  
709 used a particular exploitation or exploration strategy earlier or later in their reversals. Learning theory on  
710 serial reversal experiments predicts that all individuals in the manipulated group shifted toward the “win-  
711 stay, lose-shift” rule because their reversal speeds improved (Spence, 1936; J. Warren, 1965; J. M. Warren,  
712 1965). In contrast, learning theory on multi-armed bandit (a paradigm often used in reversal learning)  
713 decision making has a stricter criterion, predicting that the optimal strategy is to maximize the cumulative  
714 reward, which, in this case would result in individuals using the epsilon-first learning strategy immediately  
715 after the first trial (McInerney, 2010). Both learning theories consider one trial learning the optimal solution.  
716 Perhaps these wild-caught grackles relied solely on the epsilon-decreasing strategy because these individuals  
717 are used to an environment where information about the probability of what the optimal options are varies  
718 (McInerney, 2010). Therefore, maximizing information gain via continued exploration of the available options  
719 is likely of more use in the less predictable environment in the wild. Other investigations of the exploitation  
720 vs. exploration learning strategies involved in reversal learning have found that these strategies can vary by  
721 individual and relate to differences in reversal performance. For example, urban common mynas were slower  
722 to reverse a preference than rural mynas because they spent more time exploring their options (Federspiel et  
723 al., 2017). Perhaps we found no such differences in the grackles because all of the individuals we tested came  
724 from an urban area. If a rural population of grackles could be found, it would be interesting to compare  
725 learning strategy use between rural and urban individuals.

#### 726 **Why did performance on a touchscreen vary so drastically from a traditional approach?**

727 We assumed that reversal learning performance using **shape on the touchscreen** would directly compare  
728 to and be interchangeable with reversal learning performance using shaded tubes. However, it quickly  
729 became clear that the touchscreen experiment may have been asking a different question compared with  
730 the traditional reversal learning approach using physical objects. Unfortunately, we did not have the time  
731 to explore what might have caused the differences between the two tests, but we speculate below. We  
732 conclude that these two methods, the traditional physical object and the touchscreen, do not measure the  
733 same construct in this species and with this reversal learning experiment.

734 One possible explanation for the difference between the two experiments is that grackles might require more  
735 trials to learn to discriminate between shapes than between shades. Shapes are known to require a few more  
736 trials for a preference to develop (e.g., Shaw et al., 2015: mean=40 trials shade, mean=55 trials shape in  
737 toutouwai; Isden et al., 2013: mean=6 trials shade, mean=10 trials shape in spotted bowerbirds), however  
738 grackles required hundreds more trials to learn shapes, therefore this explanation seems unlikely. Moreover,  
739 grackles may not have understood how the touchscreen worked and therefore it was the apparatus that  
740 interfered with their performance, yet grackles successfully completed a go no-go inhibition task using the  
741 same touchscreen apparatus (Logan et al., 2021). The go no-go task similarly used two different white  
742 shapes (wavy lines or a heart), but the shapes were presented sequentially rather than simultaneously (as  
743 in the reversal touchscreen experiment). Given this difference between the two touchscreen experiments, it  
744 is possible that the grackles found touching the screen in the reversal experiment rewarding in and of itself  
745 because something happened whenever they made a response. That is, if they touched the correct stimulus,  
746 they received food; if they touched the incorrect stimulus, the screen went blank immediately. This is in  
747 contrast with the go no-go experiment where the stimulus stayed on the screen for a set amount of time after  
748 an incorrect choice. Another potential reason for the difference between performances on the two touchscreen  
749 experiments was that making the incorrect choice in the reversal experiment was not costly enough. In the  
750 reversal touchscreen experiment, they could get through many trials, receiving some rewards, in a short  
751 amount of time. Consequently, there was potentially not enough incentive to learn quickly, thus explaining  
752 the differences in learning speeds between the two reversal experiments.

753 We are not the first group to attempt to transfer a traditional lab or field task to a touchscreen apparatus

754 (e.g., Drayton & Santos, 2014). Despite some of the challenges associated with touchscreen apparatuses,  
755 other attempts to transfer tasks to a touchscreen have been more successful (e.g., Blaisdell & Cook, 2005;  
756 Kangas & Bergman, 2017; Sawa et al., 2005). We maintain that touchscreens have the potential to be an  
757 incredibly useful tool for studying comparative cognition in some systems (for reviews and methods, see  
758 Bussey et al., 2008; Cook et al., 2004; Kangas & Bergman, 2017; Logan et al., 2021; Seitz et al., 2021; Wolf  
759 et al., 2014).

## 760 **Conclusion**

761 We demonstrate that it is possible to manipulate flexibility, using a paradigm such as reversal learning, to  
762 examine its direct link with other traits. This opens up many opportunities for future research to better  
763 understand what flexibility is and whether and how it is causally related to other behaviors or forms of  
764 cognition. Understanding how flexibility causally relates to other traits will allow researchers to develop  
765 robust theory about the mechanisms and functional impact of flexibility, and when to invoke it as a primary  
766 driver in a given context, such as a rapid geographic range expansion. Indeed, we are already in the process  
767 of testing the latter hypothesis by conducting cross-population research on great-tailed grackles to test  
768 whether a population on the range edge is more flexible (Logan CJ et al., 2020). That we were able to  
769 manipulate flexibility, which had causal effects on flexible behavior in a different context (multi-access box)  
770 as well as a different cognitive ability (innovativeness), demonstrates that flexibility manipulations could  
771 be useful in training individuals of other species in how to be more flexible. This could have important  
772 implications for threatened and endangered taxa (such as informing the choice of individuals for captive  
773 breeding or introduction programs where individuals or their offspring are released into novel areas), as well  
774 as for habituating zoo animals or other managed populations to novelty. If such a flexibility manipulation  
775 was successful, it could then change their behavior in this and other domains, giving them a better chance of  
776 succeeding in human modified environments. This is the focus of our new research program, ManyIndividuals,  
777 where we manipulate flexibility using serial reversals in the wild in species that are successful and at risk  
778 and determine whether the manipulation improves their success in human modified environments (Logan et  
779 al., 2022).

## 780 **ETHICS**

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

- 782 1) US Fish and Wildlife Service (scientific collecting permit number MB76700A-0,1,2)
- 783 2) US Geological Survey Bird Banding Laboratory (federal bird banding permit number 23872)
- 784 3) Arizona Game and Fish Department (scientific collecting license number SP594338 [2017], SP606267  
785 [2018], and SP639866 [2019])
- 786 4) California Department of Fish and Wildlife (scientific collecting permit number S-192100001-19210-001)
- 787 5) Institutional Animal Care and Use Committee at Arizona State University (protocol number 17-1594R)
- 788 6) Institutional Animal Care and Use Committee at the University of California Santa Barbara (protocol  
789 number 958)
- 790 7) University of Cambridge ethical review process (non-regulated use of animals in scientific procedures:  
791 zoo4/17 [2017])

## 792 **AUTHOR CONTRIBUTIONS**

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

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

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



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

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

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

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

803 **McCune:** Added MAB log experiment, protocol development, data collection, data interpretation, revis-  
804 ing/editing, materials.

## 805 **FUNDING**

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

## 809 **CONFLICT OF INTEREST DISCLOSURE**

810 We, the authors, declare that we have no financial conflicts of interest with the content of this article. CJ  
811 Logan and D Lukas are Recommenders at PCI Ecology, and Logan used to be on the Managing Board  
812 (2018-2022).

## 813 **ACKNOWLEDGEMENTS**

814 We thank our PCI Ecology recommender, Aurelie Coulon, and reviewers, Maxime Dahirel, Andrea Griffin,  
815 and Aliza le Roux for their feedback on the preregistration and post-study manuscripts; Kevin Langergraber  
816 for serving as our ASU IACUC PI; Ben Trumble and Angela Bond for logistical support; Melissa Wilson  
817 for sponsoring our affiliations at Arizona State University and lending lab equipment; Kristine Johnson  
818 for technical advice on great-tailed grackles; Arizona State University School of Life Sciences Department  
819 Animal Care and Technologies for providing space for our aviaries and for their excellent support of our  
820 daily activities; Julia Cissewski for tirelessly solving problems involving financial transactions and contracts;  
821 Sophie Kaube for logistical support; Richard McElreath for project support; Aaron Blackwell and Ken Kosik  
822 for being the UCSB sponsors of the Cooperation Agreement with the Max Planck Institute for Evolution-  
823 ary Anthropology; Tiana Lam, Anja Becker, and Brynna Hood for interobserver reliability video coding;  
824 Sawyer Lung for field support; Alexis Breen for coding multi-access box videos; and our research assistants:  
825 Aelin Mayer, Nancy Rodriguez, Brianna Thomas, Aldora Messinger, Elysia Mamola, Michael Guillen, Rita  
826 Barakat, Adriana Boderash, Olateju Ojekunle, August Sevchik, Justin Huynh, Jennifer Berens, Amanda  
827 Overholt, Michael Pickett, Sam Munoz, Sam Bowser, Emily Blackwell, Kaylee Delcid, Sofija Savic, Brynna  
828 Hood, Sierra Planck, and Elise Lange.

## 829 SUPPLEMENTARY MATERIAL 1: Ability to detect actual effects

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

842 To address the power analysis issues, we ran simulations on our Arizona data set before conducting any  
843 analyses in this preregistration.

844 **Planned:** We will first run null models (i.e., dependent variable  $\sim 1 +$  random effects), which will allow us  
845 to determine what a weak versus a strong effect is for each model. Then we will run simulations based on  
846 the null model to explore the boundaries of influences (e.g., sample size) on our ability to detect effects of  
847 interest of varying strengths. If simulation results indicate that our Arizona sample size is not larger than  
848 the lower boundary, we will continue these experiments at the next field site until we meet the minimum  
849 suggested sample size.

- 850 • **Implementation of the plan:** Simulations were conducted in April 2020 (pre-data analysis) following  
851 procedures in McElreath (2018). This meant that there were no null models because the simulations  
852 using the full models are used to determine whether one can detect differences between effect sizes.

853 We first constructed a **hypothesis-appropriate mathematical model** to identify the parameter bound-  
854 aries (beta, sigma) that produce simulated data within the range of values expected for this species in the  
855 reversal learning and multi-access box experiments. Values for reversal learning using color tubes (mean,  
856 standard deviation, and range of number of trials to reverse a color preference) were taken from previously  
857 published data on great-tailed grackles (Logan, 2016). We were unsure of whether the grackles would be able  
858 to solve any options on the multi-access box because this experiment had never been done on this species  
859 before, so, in the simulation (described next), we ran versions where grackles solved between 0 and 4 options  
860 and other versions where they solved between 0 and 2 options (out of 4 options maximum). The model is  
861 as follows:

$$862 t_{i,j} \sim \text{Normal}(\mu, \sigma) \text{ [likelihood]},$$

$$863 \mu \sim \alpha + \beta x \text{ [linear model]},$$

$$864 \alpha \sim \text{Normal}(91,21) \text{ [\alpha prior]},$$

$$865 \beta \sim \text{Normal}(0,0.5) \text{ [\beta prior]},$$

$$866 \sigma \sim \text{Uniform}(0,40) \text{ [\sigma prior]},$$

867 where  $t_{i,j}$  is the number of trials to reverse a preference (with fewer trials indicating faster reversal and thus  
868 more flexibility),  $\mu$  is the population average trials to reverse,  $\sigma$  is the population standard deviation for  
869 trials to reverse,  $\alpha$  is the intercept,  $\beta$  is the expected amount of change in  $t_{i,j}$  for every one unit change in  
870  $x$ , and  $x$  is the number of options solved on the multi-access box. We used a normal distribution for  $t_{i,j}$ ,  $\alpha$ ,  
871 and  $\beta$  because they are (or are based on) sums with large means (see Figure 10.6 in McElreath, 2016). We  
872 plugged in data from the Santa Barbara grackles (Logan, 2016): 91=average number of trials to reverse a  
873 preference (standard deviation=21 trials). The  $\beta$  prior uses 0 as the mean and 0.5 as the standard deviation  
874 as a place to start and may need to be adjusted as the data are simulated. We chose a uniform distribution  
875 for  $\sigma$  because it constrains  $\sigma$  to have a positive probability of the standard deviation being between 0 and 40

876 trials (range of the number of trials to reverse a preference: 39-130, therefore 130 trials minus the average  
 877 number of trials (91), which is about 40).

878 We **translated the simulation output into effect sizes** and examined what kind of effect size we could  
 879 detect (Table SM1.1). For each  $\beta$ , we calculated the effect size as in Lajeunesse et al. (2013; Box 13.3 in  
 880 Lajeunesse et al., 2013: linear regression):

$$881 \quad r = \beta (SDx_{i,j} / SDy_{i,j}) = \beta (1.5 / 21),$$

882 where  $r$  is the Pearson product moment correlation and  $SD$  is the standard deviation. For the standard  
 883 deviation of  $x_{i,j}$  (number of loci solved on the multiaccess box), we estimated a possible value of 1.5. For  
 884 the standard deviation of  $y_{i,j}$  (trials to reverse), we used 21 from the Santa Barbara grackle data (Logan,  
 885 2016). We then calculated the effect sizes and  $R^2$  values for each value of  $\beta$ .

886 **Table SM1.1.** The connection between  $\beta$  and effect sizes ( $SDx_{i,j}$ =standard deviation of  $x_{i,j}$ , which is the  
 887 number of loci solved;  $SDy_{i,j}$ =standard deviation of  $y_{i,j}$ , which is the number of trials to reverse).

	Beta	SDx	SDy	Effect size	R-squared
888	-5	1.5	21	-0.357	0.128
	-1	1.5	21	-0.071	0.005
889	0	1.5	21	0.000	0.000

890 We then used the simulations to run **models** on simulated data to estimate the measurement error associated  
 891 with varying sample size,  $\beta$ , and the range of multi-access box loci solved or latency to attempt a new locus  
 892 (Table SM1.2). Before running the models, we decided that a model would detect an effect if 89% of the  
 893 posterior sample was on the same side of zero (following McElreath, 2018). We ran the simulation with  
 894  $\beta=5$  because this was a high value at which an appropriate range of values were observed in the simulation  
 895 testing phase,  $\beta=0$  because this would be the scenario in which there is no relationship between the response  
 896 variable and the trials to reverse, and  $\beta=-1$  to determine how small of a difference we can detect and with  
 897 what amount of associated noise ( $\sigma$ ). Sigma ( $\sigma$ ) is the standard deviation in the trials to reverse if the  
 898 trials to reverse is a normal distribution. In all simulations, the mean in the trials to reverse was set to 91.  
 899 Therefore, a ( $\sigma$ ) of 14 is 15% noise (14/91). We found that when ( $\sigma$ ) is larger than 14, we cannot detect  
 900 even the largest effect of trials to reverse on loci solved or latency because there are some simulations where  
 901 the estimated regression coefficient crosses zero. When  $\beta=0$  we want all of the regression coefficients to cross  
 902 zero (10 out of 10 random repetitions) and when  $\beta \neq 0$  we want none of the regression coefficients to cross  
 903 zero (0 out of 10 random repetitions). We ran the models several times with various parameters to determine  
 904 at what point this was the case for each combination of parameters.

905 **Table SM1.2.** Simulation outputs from varying  $\beta$ , sample size ( $n$ ),  $\sigma$ , and whether the actual range of  
 906 multi-access box [MAB] loci solved were 0-2 or 0-4 (we did not know how many loci the grackles would be  
 907 able to solve before we started collecting data so we ran two simulations. The grackles ended up being able to  
 908 solve all four loci on both multi-access boxes, therefore we must use only those rows associated with “Range  
 909 of MAB loci solved” = 0-4). We ran the simulation with  $\beta$  at -5 because this was what ended up generating  
 910 an appropriate range of values in the parameter testing phase, at 0 because this would be the scenario in  
 911 which there is no relationship between trials to reverse and number of multi-access box loci solved, and -1  
 912 to determine how small of a difference we can detect with what amount of associated error ( $\sigma$ ). When  $\beta$   
 913 = 0 we want all of the regression coefficients to cross zero (10/10) and when  $\beta \neq 0$  we want none of the  
 914 regression coefficients to cross zero (0/10). We used the simulations to determine at what point this was the  
 915 case for each combination of parameters. This table is useful for the analyses involving the number of loci  
 916 solved on the multi-access box, but not the latency to switch to attempting a new locus on the multi-access  
 917 box, which uses a different (gamma poisson) model.

Beta	n	Sigma	Regression coefficient crosses zero	Regression coefficient	Range of MAB loci solved
-5	15	15	1/10	-5.90	0-4
-5	15	14	0/10	-5.11	0-4
-5	15	12	0/10	-4.79	0-4
-5	15	10	0/10	-4.31	0-4
-5	10	10	1/10	-4.35	0-4
-5	10	9	0/10	-5.26	0-4
-5	8	10	1/10	-5.35	0-4
-5	8	9	0/10	-4.22	0-4
-5	8	8	0/10	-3.08	0-4
-5	8	8	1/10	-4.74	0-2
-5	8	7	3/10	-6.74	0-2
-5	8	5	0/10	-3.08	0-2
-5	10	9	3/10	-4.51	0-2
-5	10	7	1/10	-7.67	0-2
-5	10	6	2/10	-5.16	0-2
-5	10	5	1/10	-4.57	0-2
-5	10	4	0/10	-5.02	0-2
-5	15	14	2/10	-3.07	0-2
-5	15	13	5/10	1.68	0-2
-5	15	10	5/10	-8.20	0-2
-5	15	8	3/10	-4.01	0-2
-5	15	6	0/10	-6.03	0-2
-5	15	7	1/10	-8.06	0-2
0	15	14	10/10	-3.23	0-2
0	15	14	10/10	0.43	0-4
-1	15	14	10/10	-1.53	0-4
-1	15	10	10/10	-0.73	0-4
-1	15	5	3/10	0.19	0-4
-1	15	3	1/10	0.18	0-4
-1	15	2	0/10	-1.07	0-4
-1	15	2	3/10	-1.67	0-2
-1	15	1	1/10	-1.12	0-2

918

919

920 This shows that we would have the power to detect a medium effect (-0.357 in Table SM1.1) with a sample  
921 size of 15 if the noise ( $\sigma$ ) is <15%. We would be unlikely to get a false negative because there were no false  
922 negatives in the simulations (i.e., the posterior sample range did not cross zero). With this sample size, when  
923  $\beta=0$ , there are no false positives (i.e., the posterior sample range always included zero). However, we would  
924 not be able to detect a weak effect unless the noise ( $\sigma$ ) was much smaller.

### 925 **Simulation and model: number of loci solved on the multi-access box ~ trials to reverse**

926 The model takes the form of:

$$927 \text{locisolved}_{i,j} \sim \text{Binomial}(4, p) \text{ [likelihood]},$$

$$928 \text{logit}(p) \sim \alpha[\text{batch}] + \beta \text{trials}_{i,j} \text{ [model]},$$

929 where  $\text{locisolved}_{i,j}$  is the number of loci solved on the multi-access box, 4 is the total number of loci on the  
930 multi-access box,  $p$  is the probability of solving any one locus across the whole experiment,  $\alpha$  is the intercept  
931 and each batch gets its own,  $\beta$  is the expected amount of change in  $\text{locisolved}_{i,j}$  for every one unit change  
932 in  $\text{trials}_{i,j}$ , and  $\text{trials}_{i,j}$  is the number of trials to reverse a shade preference.

933 Expected values for the number of loci solved on the multi-access box were set to either 2 or 0 (out of  
934 4 loci maximum) because we were unsure of whether the grackles would be able to solve any loci on the  
935 multi-access box because this experiment had never been done on this species before. Expected values for  
936 reversal learning using shaded tubes (mean, standard deviation, and range of number of trials to reverse a  
937 shade preference) were based on previously published data on great-tailed grackles (Logan, 2016). This data  
938 indicates that the average number of trials to reverse a preference is 91 and the standard deviation is 21. In

939 our model, the variation in the actual data is reflected by both the population standard deviation and the  
940 expected amount of change related to the explanatory variable. After running simulations, we identified the  
941 following distributions and priors to be the most likely for our expected data:

$$942 \alpha \sim \text{Normal}(4,10) [\alpha \text{ prior}],$$

$$943 \beta \sim \text{Normal}(0,5) [\beta \text{ prior}].$$

944 We used normal distributions for  $\alpha$  and  $\beta$  because they are (or are based on) sums with large means (see  
945 Figure 10.6 in McElreath, 2018). For the  $\beta$  prior, we had no expectation about whether the relationship  
946 would be positive or negative, therefore we centered it on 0 (the mean).

#### 947 **Simulation and model: latency to attempt a new locus on the multi-access box ~ trials to** 948 **reverse**

949 For the average latency to attempt a new locus on the multi-access box as it relates to trials to reverse (both  
950 are measures of flexibility), we simulated data and set the model as follows:

$$951 \textit{latency}_{i,j} \sim \text{gamma-Poisson}(\lambda_{i,j}, \phi) [\textit{likelihood}],$$

$$952 \log(\lambda_{i,j}) \sim \alpha[\textit{batch}] + \beta \textit{trials}_{i,j} [\textit{model}],$$

953 where  $\textit{latency}_{i,j}$  is the average latency to attempt a new locus on the multi-access box,  $\lambda_i$  is the random  
954 probability of attempting a locus in each second per bird (and we take the log of it to make sure it is always  
955 positive; birds with a higher rate have a smaller latency),  $\phi$  is the dispersion of the rates across birds,  $\alpha$  is  
956 the intercept for the rate per batch,  $\beta$  is the expected amount of change in the rate of attempting to solve  
957 in any given second for every one unit change in  $\textit{trials}_{i,j}$ , and  $\textit{trials}_{i,j}$  is the number of trials to reverse a  
958 shade preference.

959 Expected values for the latency to attempt a new locus on the multi-access box was set to between 1-2700  
960 sec because the experiment ends for a bird if they do not obtain the food in 3 consecutive trials, and each  
961 trial can last up to 15 min (trials end at 10 min unless the individual is on the ground at the 10 min mark,  
962 in which case they are given an extra 5 min to interact). Because we did not have prior data for this species  
963 on this test, we set the mean to 300 sec, which is half way through a usual 10 min trial because it seems  
964 likely that if a bird is going to attempt another locus, it will likely do so at the next opportunity, especially  
965 after being successful in the previous trial. Expected values for reversal learning using shaded tubes are the  
966 same as above. After running simulations, we identified the following to be the most likely distributions and  
967 priors for our expected data:

$$968 \phi \sim 1/\textit{exp}(1) [\phi \text{ prior}],$$

$$969 \alpha \sim \text{Normal}(300,50) [\alpha \text{ prior}],$$

$$970 \beta \sim \text{Normal}(0,5) [\beta \text{ prior}].$$

971 We used a gamma-Poisson distribution for  $\textit{latency}_{i,j}$  because it constrains the values to be positive and to  
972 primarily occur sooner rather than later, which is what we expect from the grackles (based on data from New  
973 Caledonian crows and kea in Auersperg et al., 2011). For  $\phi$ , we used an exponential distribution because it  
974 is standard for this parameter. We used normal distributions for  $\alpha$  and  $\beta$  because they are (or are based on)  
975 sums with large means (see Figure 10.6 in McElreath, 2018). For the  $\beta$  prior, we had no expectation about  
976 whether the relationship would be positive or negative, therefore we centered it on 0 (the mean).

977 **SUPPLEMENTARY MATERIAL 2: Interobserver reliability of dependent vari-**  
978 **ables (unregistered analyses)**

979 To determine whether experimenters coded the dependent variables in a repeatable way, hypothesis-blind  
980 video coders were first trained in video coding the dependent variable, and then they coded at least 20% of  
981 the videos in the reversal (tubes) and multi-access box experiments. We randomly chose a subset of all of  
982 the birds who participated in each experiment using random.org:

- 983 • Reversal 6/20 grackles (30% with half from the control group): Chalupa, Avocada, Diablo, Fideo,  
984 Tomatillo, Adobo
- 985 • Multi-access box plastic 3/15 grackles (20%): Habanero, Queso, Chalupa
- 986 • Multi-access box log 3/12 grackles (25%): Diablo, Adobo, Yuca

987 Video coders then analyzed all videos from these birds. The experimenter's data was compared with the  
988 video coder data using the intra-class correlation coefficient (ICC) to determine the degree of bias in the  
989 regression slope (Hutcheon et al. (2010), using the irr package in R: Gamer et al. (2012)). Note that the  
990 data in columns from coders 1 and 2 in the data sheets were aligned based on similar numbers between  
991 coders to prevent disagreements near the top of the data sheet from misaligning all subsequent entries.

992 **INTEROBSERVER RELIABILITY TRAINING**

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

996 *Alexis Breen* (compared with experimenter's live coding):

- 997 • Multi-access box: correct choice unweighted Cohen's Kappa=0.90 (confidence boundaries=0.77-1.00,  
998 n=33 data points)
- 999 • Multi-access box: locus solved unweighted Cohen's Kappa=0.90 (confidence boundaries=0.76-1.00,  
1000 n=33 data points)

1001 Note: Breen was not a hypothesis-blind video coder. She contributed to extensive video coding across  
1002 the whole project, however, for interobserver reliability analyses, her data were always compared with a  
1003 hypothesis-blind coder's data.

1004 *Anja Becker* (compared with experimenter's live coding):

- 1005 • Reversal: correct choice ICC=1.00 (confidence boundaries=1.00-1.00, n=25 data points)

1006 *Tiana Lam* (compared with experimenter's live coding):

- 1007 • Multi-access box: correct choice ICC=0.90 (confidence boundaries=0.77-1.00, n=33 data points)
- 1008 • Multi-access box: locus solved unweighted Cohen's Kappa=0.95 (confidence boundaries=0.84-1.00,  
1009 n=33 data points)

1010 *Brynna Hood* (compared with experimenter's live coding):

- 1011 • Multi-access log: correct choice unweighted Cohen's Kappa=1.00 (confidence boundaries=1.00-1.00,  
1012 n=29 data points)
- 1013 • Multi-access log: locus solved unweighted Cohen's Kappa=1.00 (confidence boundaries=1.00-1.00,  
1014 n=29 data points)

1015 **INTEROBSERVER RELIABILITY**

1016 Interobserver reliability scores (minimum 20% of the videos) were as follows:

1017 ***Brynna Hood*** (compared with experimenter's live coding):

- 1018 • Multi-access log: correct choice unweighted Cohen's Kappa=0.91 (confidence boundaries=0.76-1.00,  
1019 n=39 data points)
- 1020 • Multi-access log: locus solved unweighted Cohen's Kappa=1.0 (confidence boundaries=1.0-1.00, n=39  
1021 data points)

1022 ***Tiana Lam*** (compared with experimenter's live coding):

- 1023 • Multi-access box: correct choice unweighted Cohen's Kappa=0.83 (confidence boundaries=0.73-0.92,  
1024 n=102 data points)
- 1025 • Multi-access box: locus solved unweighted Cohen's Kappa=0.90 (confidence boundaries=0.830-0.97,  
1026 n=102 data points)

1027 ***Anja Becker*** (compared with experimenter's live coding):

- 1028 • Reversal: correct choice ICC=0.99 (confidence boundaries=0.98-0.99, n=3280 data points)

1029 These scores indicate that the dependent variables are repeatable to a high or extremely high degree given  
1030 our instructions and training

1031 **SUPPLEMENTARY MATERIAL 3: Prediction 2 model outputs**

1032 **Table SM3.** Model outputs for the number of loci solved and the latency to switch loci after passing  
1033 criterion on a different locus on the plastic (models 1-5 and 9-11) and wooden (models 6-8 and 12-14)  
1034 multi-access boxes, and for the pairwise comparisons explaining the changes caused by the manipulation  
1035 (Models 15-16). SD=standard deviation, the 89% prediction intervals are shown, n\_eff=effective sample  
1036 size, Rhat4=an indicator of model convergence (1.00 is ideal), a=the intercept (a[batch] is the intercept for  
1037 each batch), b=the slope of the relationship between loci solved or average switch latency and the number  
1038 of trials to pass the reversal. See Supplementary Material 1 for details on model specifications.



	Mean	SD	Lower 89 percentile compatibility interval (5.5%)	Upper 89 percentile compatibility interval (94.5%)	n_eff	Rhat4
MODEL 1 (last reversal): loci solved plastic ~ a[batch] + b*trials						
a[1]	0.04	0.46	-0.70	0.78	2304	1.00
a[2]	0.29	0.36	-0.30	0.87	2456	1.00
a[3]	-0.78	0.55	-1.65	0.08	2510	1.00
b	-0.22	0.25	-0.63	0.18	2364	1.00
MODEL 2 (last reversal): loci solved plastic ~ a + b*trials						
a	-0.02	0.24	-0.40	0.35	1466	1.00
b	-0.46	0.31	-0.97	-0.01	1383	1.00
MODEL 3 (last reversal): trials ~ a[batch]						
a[1]	0.09	0.37	-0.48	0.69	2095	1.00
a[2]	-0.21	0.29	-0.68	0.25	1715	1.00
a[3]	0.25	0.39	-0.38	0.86	2161	1.00
sigma	1.03	0.21	0.75	1.39	2049	1.00
MODEL 4: loci solved ~ a[condition]						
a[1] control	-0.11	0.32	-0.62	0.40	1311	1.00
a[2] manipulated	0.15	0.39	-0.46	0.80	1222	1.00
MODEL 5 (first reversal): loci solved plastic ~ a + b*trials						
a	0.00	0.24	-0.37	0.39	1208	1.00
b	-0.44	0.30	-0.94	0.02	1273	1.00
MODEL 6 (last reversal): loci solved wooden ~ a + b*trials						
a	1.06	0.27	0.63	1.50	1255	1.00
b	0.41	0.43	-0.21	1.13	1107	1.00
MODEL 7: loci solved ~ a[condition]						
a[1] control	-0.45	0.40	-1.10	0.18	1161	1.00
a[2] manipulated	0.77	0.41	0.13	1.44	1302	1.00
MODEL 8 (first reversal): loci solved wooden ~ a + b*trials						
a	0.11	0.26	-0.30	0.52	1221	1.00
b	-0.50	0.35	-1.09	0.04	1234	1.00
MODEL 9 (last reversal): avg switch latency plastic ~ a + b*trials						
a	4.93	0.30	4.45	5.41	1235	1.01
b	0.46	0.29	0.00	0.92	1363	1.00
phi	0.93	0.35	0.44	1.55	1476	1.00
MODEL 10: avg switch latency plastic ~ a[condition]						
a[1] manipulated	4.07	0.39	3.46	4.68	1027	1.00
a[2] control	5.18	0.39	4.50	5.76	1006	1.00
phi	0.91	0.41	0.37	1.63	925	1.01
MODEL 11 (first reversal): avg switch latency plastic ~ a + b*trials						
a	4.93	0.29	4.46	5.39	1488	1.00
b	0.46	0.28	0.02	0.93	1211	1.00
phi	0.94	0.36	0.44	1.60	1447	1.00
MODEL 12 (last reversal): avg switch latency wooden ~ a + b*trials						
a	5.75	0.28	5.28	6.18	1049	1.00
b	-0.41	0.32	-0.86	0.15	1281	1.01
phi	1.04	0.42	0.48	1.77	1456	1.00
MODEL 13: avg switch latency wooden ~ a[condition]						
a[1] control	5.31	0.42	4.61	5.95	701	1.00
a[2] manipulated	5.34	0.44	4.61	6.00	620	1.01
phi	0.66	0.32	0.25	1.25	806	1.00
MODEL 14 (first reversal): avg switch latency wooden ~ a + b*trials						
a	5.71	0.26	5.28	6.12	1109	1.00
b	-0.50	0.28	-0.89	-0.01	1308	1.00
phi	1.08	0.41	0.53	1.80	1347	1.00
MODEL 15 (improvement): trials ~ a[bird] + b[bird]*reversal						
b_bar	-30.30	3.51	-35.65	-24.65	109	1.00
sigma_bar	2.13	2.93	0.17	9.77	9	1.00
sigma	6.54	2.42	0.23	9.41	10	1.00
MODEL 16 (improvement): trials ~ a[reversal] + b[bird,reversal]						
rho	0.34	0.39	-0.40	0.85	2452	1.00

1039

1040

1041 **SUPPLEMENTARY MATERIAL 4: Reversal learning experiments: discrimi-**  
1042 **nating shapes on the touchscreen compared with shade using tubes**

1043 In the tube experiment, it took four grackles an average of 40 trials (sd=12) in the initial discrimination  
1044 phase to learn to prefer a shade, while it took the same individuals an average of 390 trials (sd=59) to learn  
1045 to prefer a shape using the touchscreen (Queso, Mole, Habanero, and Tapa). The two individuals who were  
1046 faster to learn in the tube experiment were slower to learn in the touchscreen experiment. For the reversal,  
1047 it took three of these individuals (Queso, Mole, and Habanero) an average of 80 trials (sd=14) to reverse  
1048 their shaded tube preference, and an average of 362 trials (sd=111) to reverse their shape preference on the  
1049 touchscreen (Tapa had to be released back to the wild before finishing the experiment, but was on trial 629  
1050 in reversal one of the touchscreen experiment at the time of release. In the tube experiment, she was also  
1051 the slowest of the four to reverse at 100 trials). All three individuals were about equally fast at the reversal  
1052 in the tube experiment, while their reversal learning speeds differed on the touchscreen. The touchscreen  
1053 training data and a summary of the training process is detailed in Seitz et al. (2021).

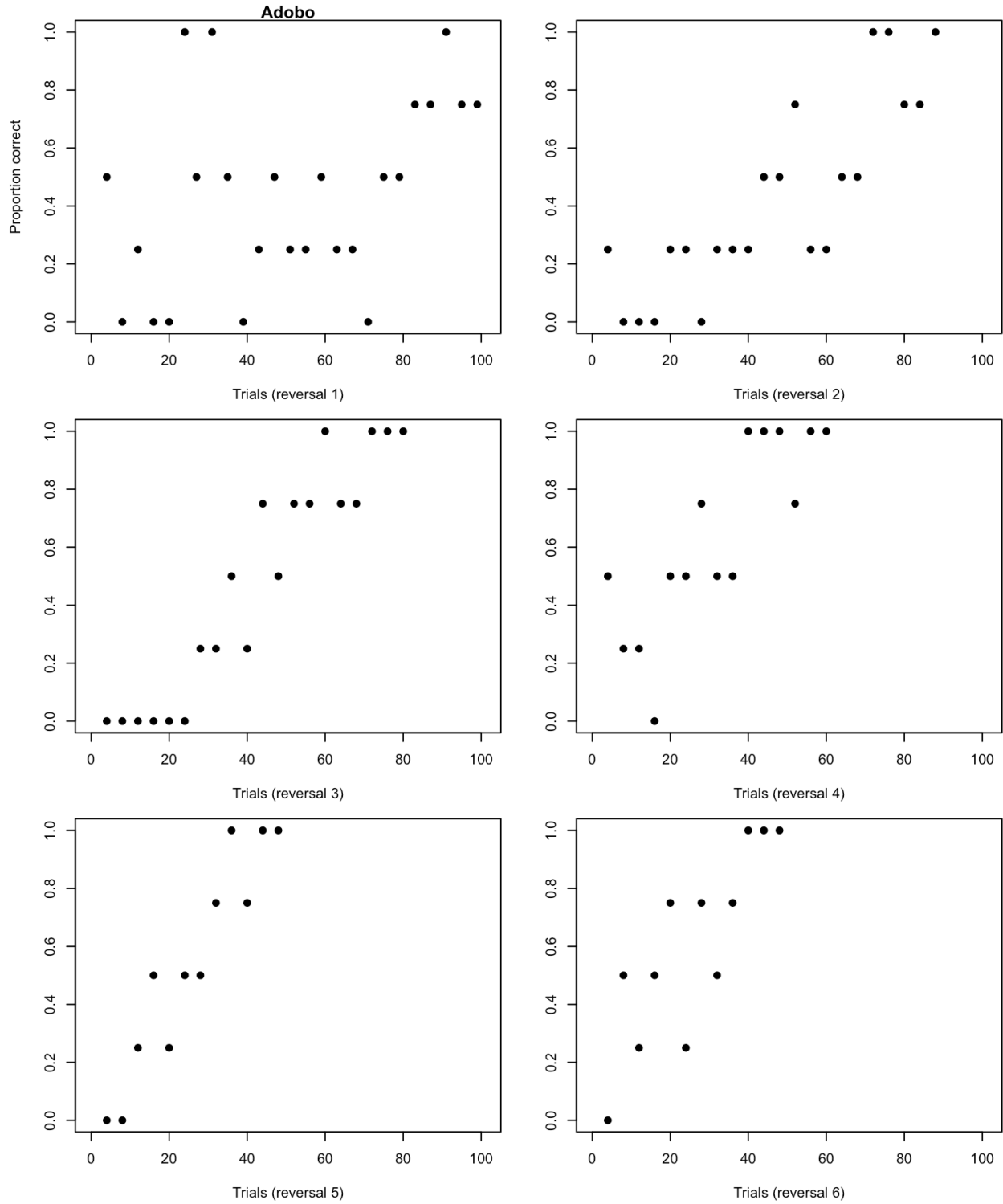


1055 **Table SM5.** Summarized results per bird in the reversal learning (tube and touchscreen) and multi-access box (plastic and wooden) experiments.  
1056 “Reversals to pass” indicates how many serial reversals it took a bird to pass criterion (passing two consecutive reversals in 50 trials or less) if they  
1057 were in the flexibility manipulation condition. X indicates the bird attempted, but did not pass that experiment. Note: Tapa did not finish the MAB  
1058 log experiment; Marisco’s MAB log experiment ended too early due to experimenter error (timed out on 2 consecutive sessions, not 3); Mole and  
1059 Habanero: do not count MAB plastic number of options solved because they were given the box fully put together for habituation due to experimenter  
1060 error; Taco was the first juvenile we tested and we did not put him in the flexibility experiment: he received 1 reversal and moved on to his next test,  
1061 therefore he was essentially a control bird without the matched yellow tube experience.

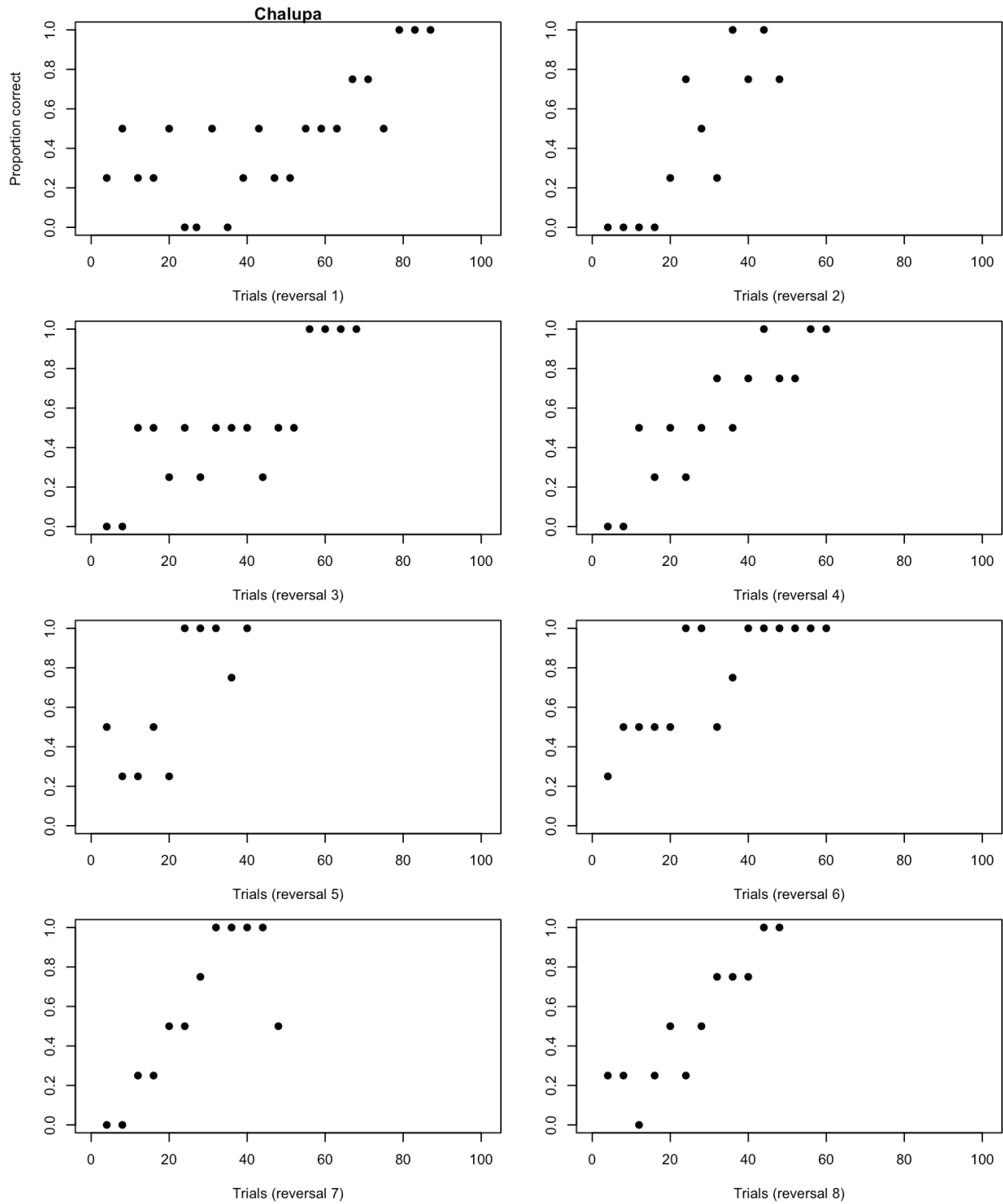
Bird	Batch	Sex	Trials to learn (tube)	Trials to first reversal (tube)	Trials to last reversal (tube)	Reversals to pass	Total loci solved (MAB plastic)	Total loci solved (MAB wooden)	Average latency to attempt new locus (MAB plastic)	Average latency to attempt new locus (MAB wooden)	Trials to learn (touchscreen)	Trials to first reversal (touchscreen)	Motor actions (MAB plastic)	Motor actions (MAB wooden)
Tomatillo	1	M	40	50	50	Control	3		317				13	
Queso	1	M	50	70	70	Control	1		88		330	460	8	
Tapa	1	F	30	100	100	Control	4		685		450	(629+)	12	
Yuca	3	F	40	80	80	Control	4	4	132	77			13	16
Marisco	3	M	40	50	50	Control	1	2		208			3	7
Pizza	3	M	50	60	60	Control	0	1		1482			0	8
Mofongo	4	M	20	40	40	Control	3	4	502	630			13	14
Taquito	4	M	90	160	160	Control	0	4		100			11	10
Chalupa	1	F	50	90	50	8	0						6	
Mole	1	M	30	70	50	7	4	4	356	1173	431	307	14	15
Habanero	1	M	50	80	40	6	4		28		350	290	15	
Diablo	3	M	20	80	40	8	2	1	25				10	2
Burrito	3	M	40	60	23	8	3	4	76	391			17	18
Adobo	3	M	50	100	50	6	4	4	31	79			16	18
Chilaquile	3	JM	30	40	30	6	4	4	44	170			19	11
Pollito	4	M	40	60	40	8	0	3		668			0	11
Taco	3a	JM	50	80	80	(Control)	1	4		117			2	19
Memela	1	F	50	60	80	X (11+)								
Fideo	2	M	60	70	70	Control								
Avocada	1	F	50	100	100	Control								
Huachinago	3	M	70			Control								
Guacamole	4	M	30											

1064 **SUPPLEMENTARY MATERIAL 6: Prediction 4 learning strategy figures**

1065 Below are figures for the proportion of trials correct by trial number and reversal for each bird.

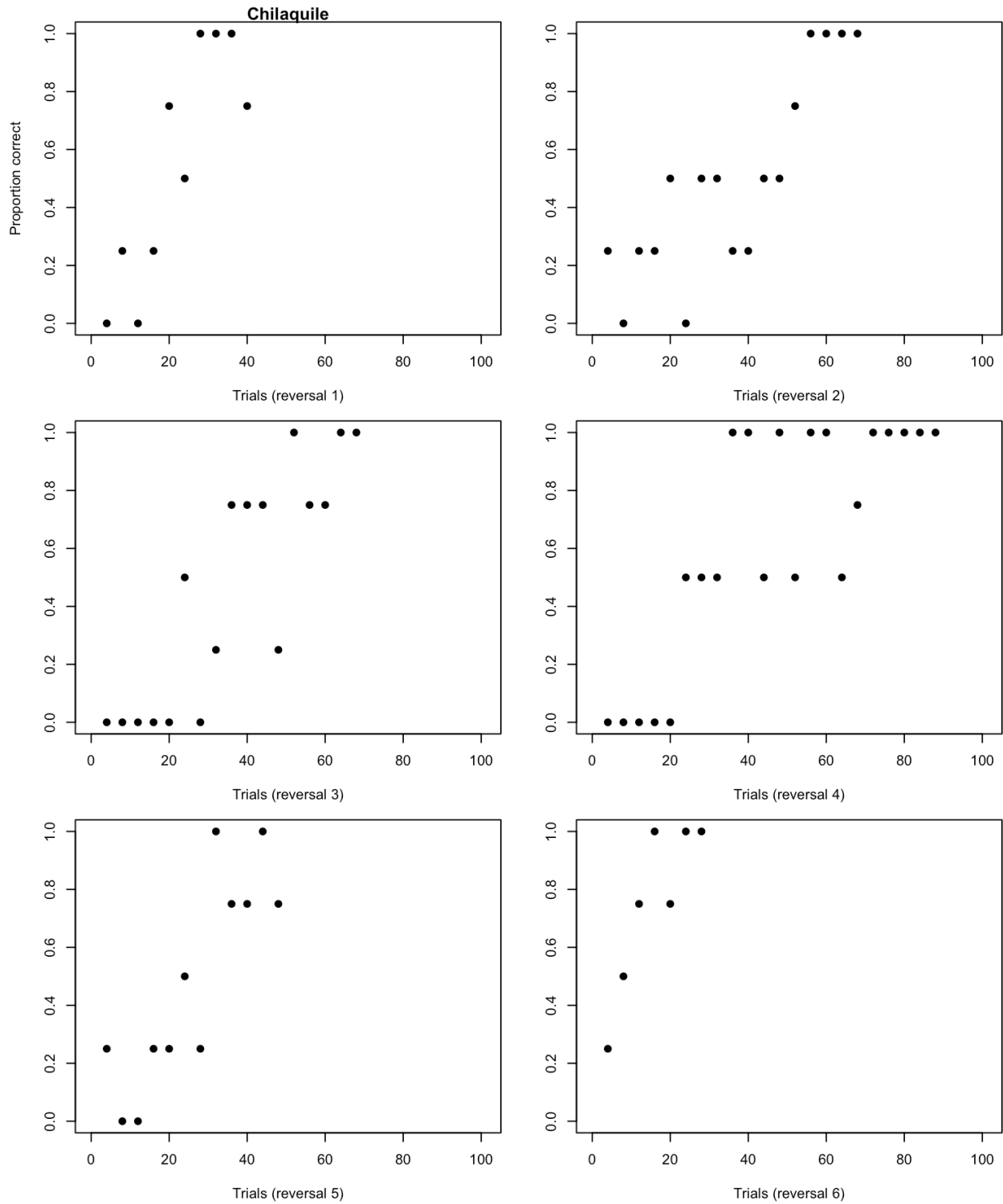


1066  
1067 **Figure SM6.1.** Adobo's proportion of trials correct by trial number and reversal.



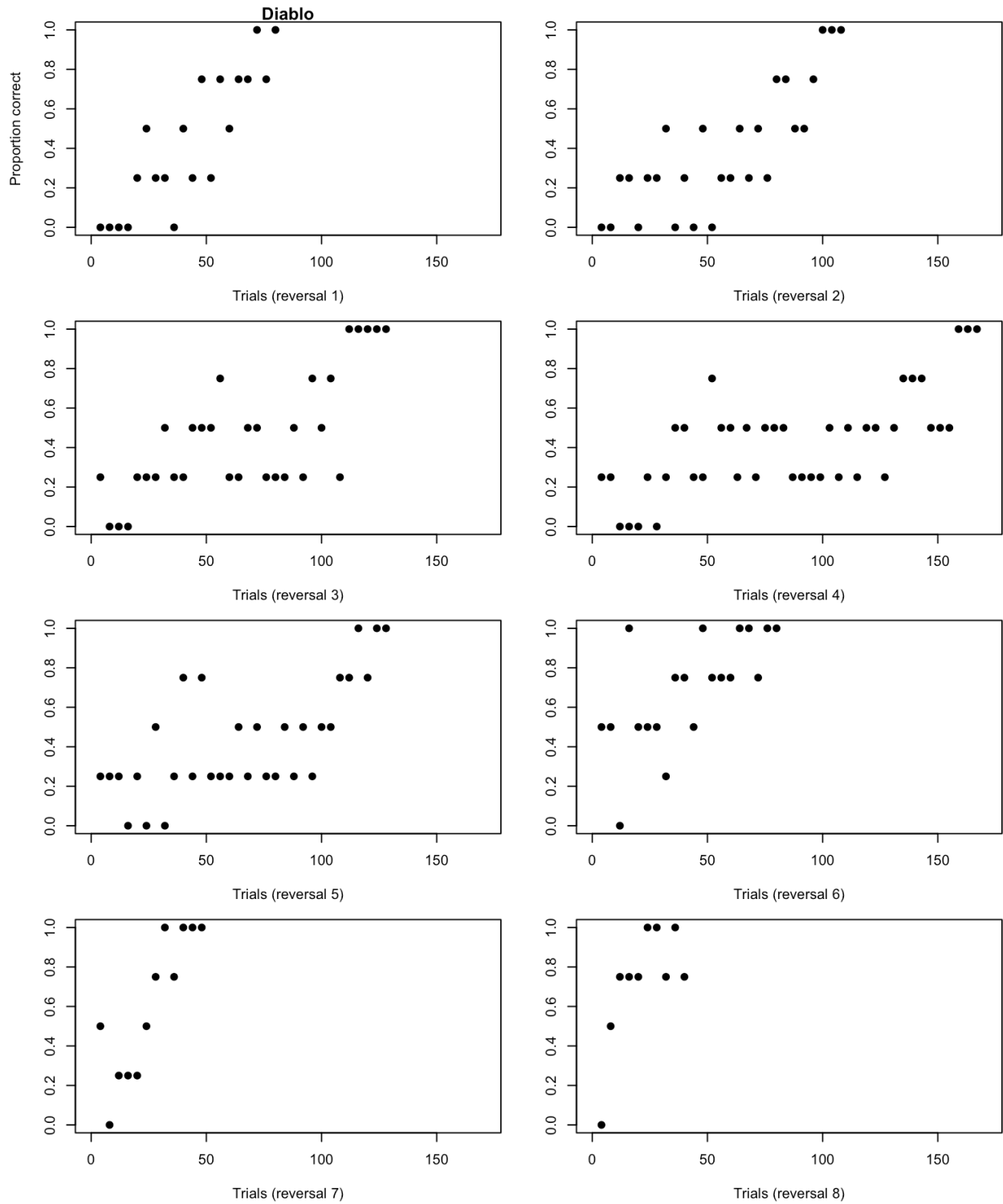
1068

1069 **Figure SM6.2.** Chalupa's proportion of trials correct by trial number and reversal.



1070

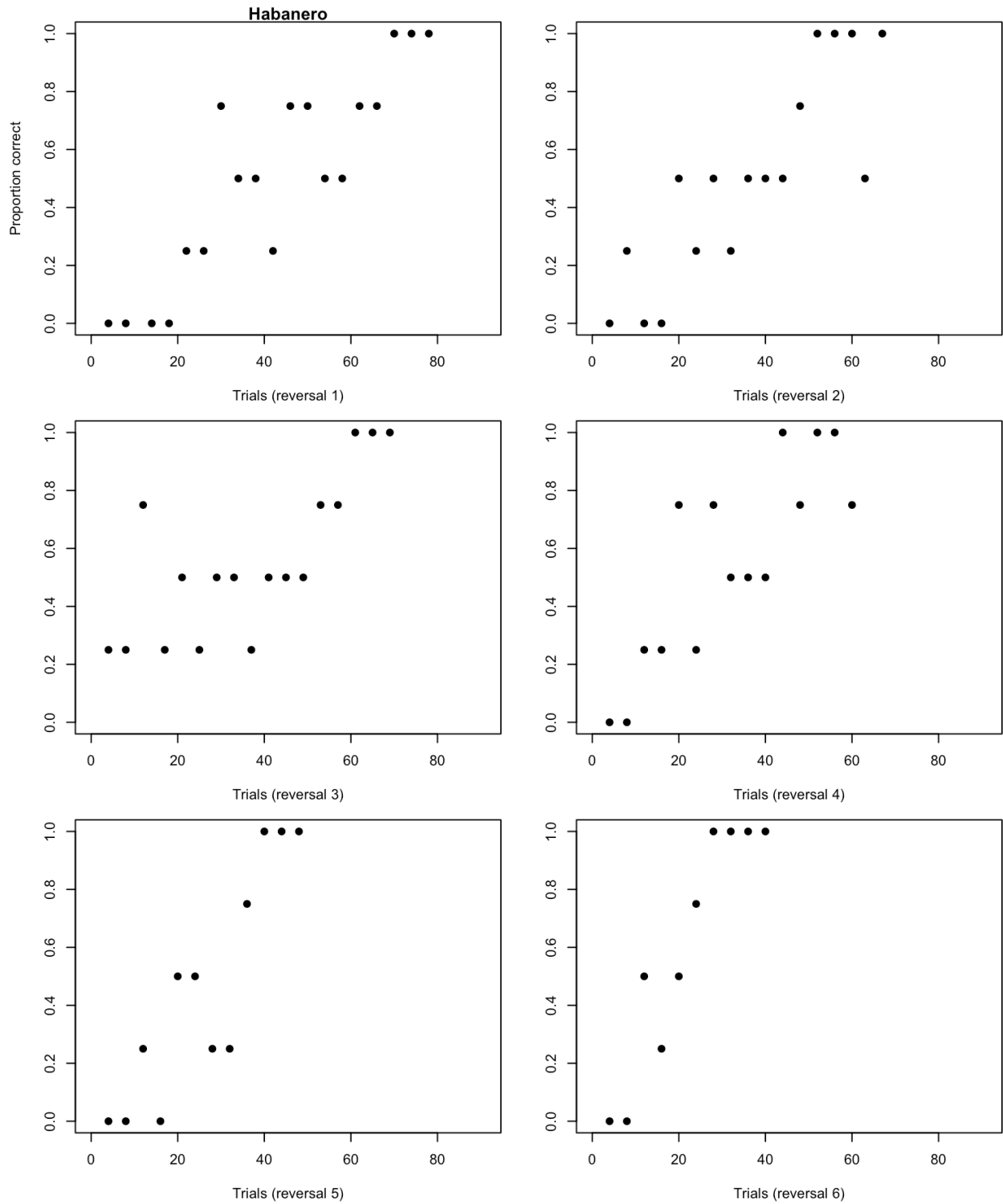
1071 **Figure SM6.3.** Chilaquile's proportion of trials correct by trial number and reversal.



1072

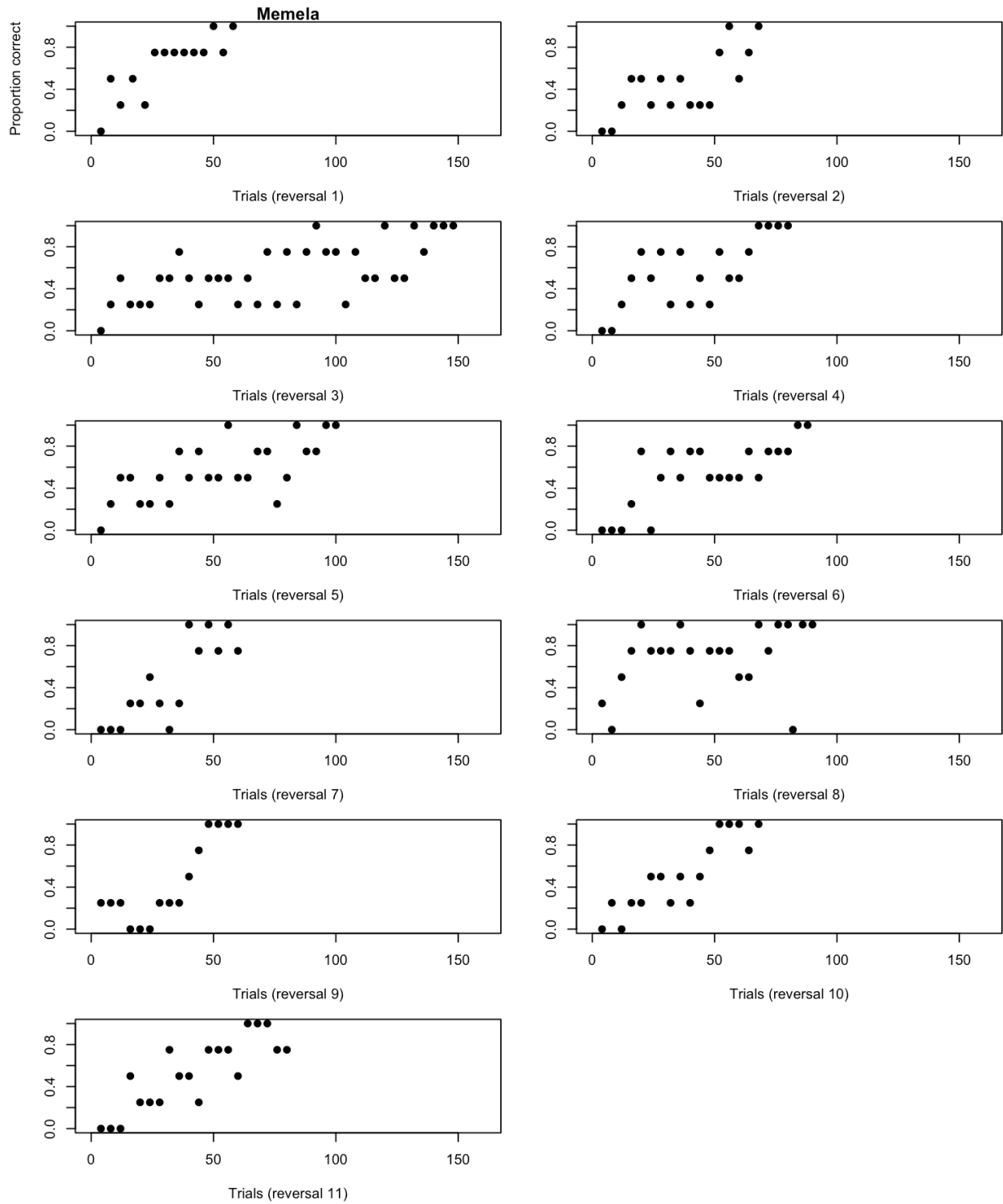
1073 **Figure SM6.4.** Diabolo's proportion of trials correct by trial number and reversal.





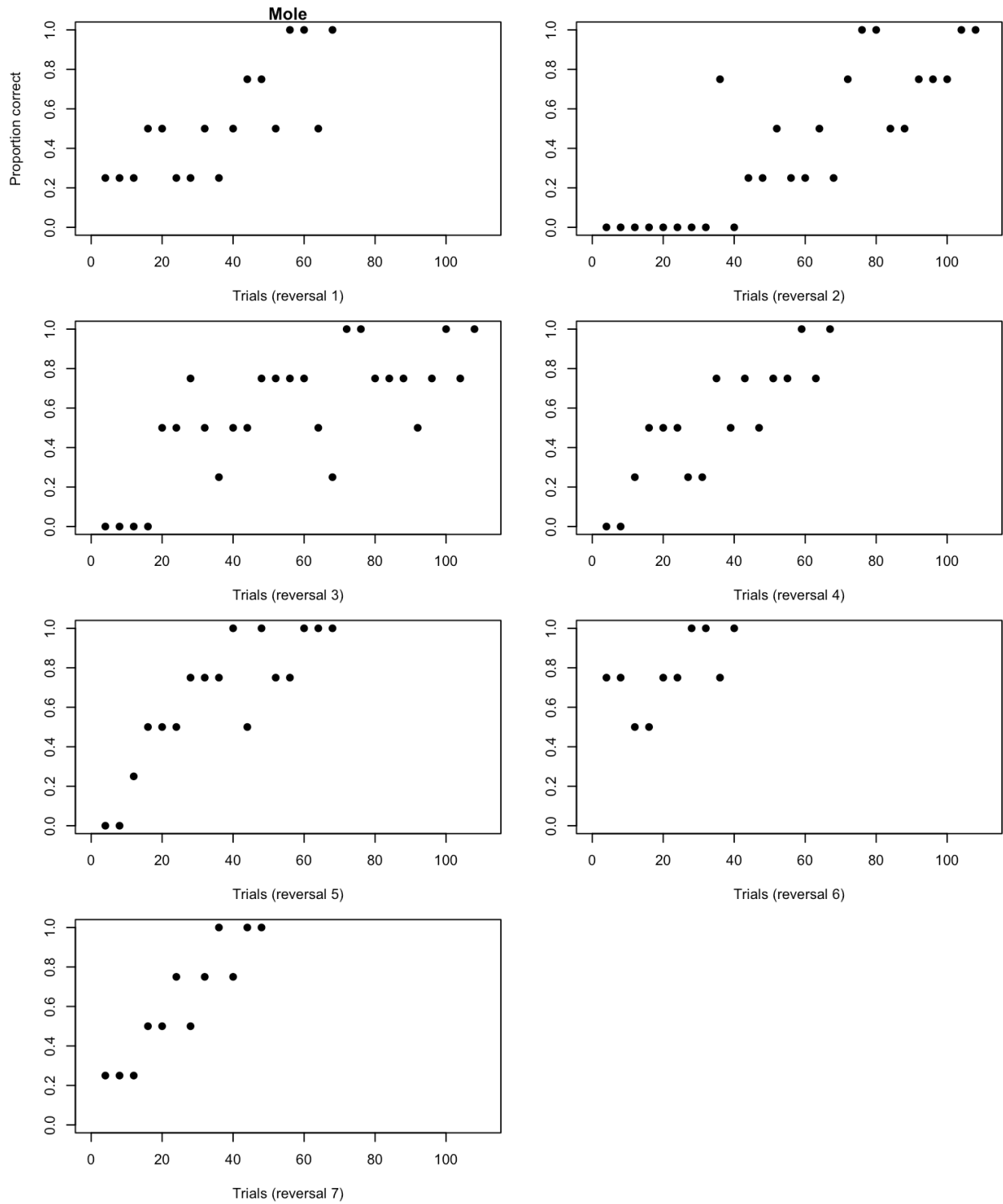
1074

1075 **Figure SM6.5.** Habanero's proportion of trials correct by trial number and reversal.



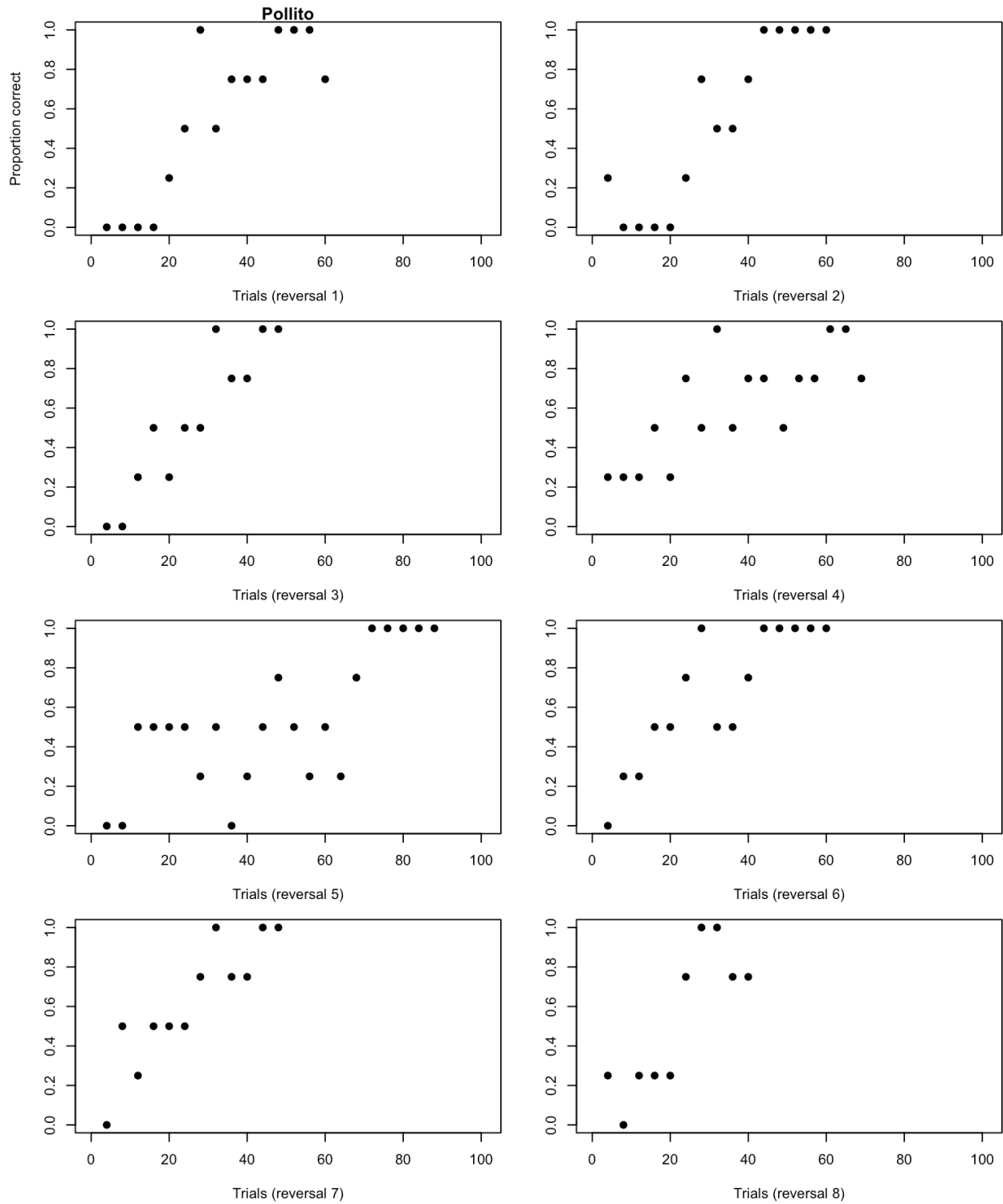
1076

1077 **Figure SM6.6.** Memela's proportion of trials correct by trial number and reversal.



1078

1079 **Figure SM6.7.** Mole's proportion of trials correct by trial number and reversal.



1080

1081 **Figure SM6.8.** Pollito's proportion of trials correct by trial number and reversal.

## REFERENCES

- 1082
- 1083 Aplin, L. M., Farine, D. R., Morand-Ferron, J., Cockburn, A., Thornton, A., & Sheldon, B. C. (2015).  
 1084 Experimentally induced innovations lead to persistent culture via conformity in wild birds. *Nature*,  
 1085 *518*(7540), 538–541.
- 1086 Auersperg, A.M.I., Bayern, A. M. P. von, Gajdon, G. K., Huber, L., & Kacelnik, A. (2011). Flexibility in  
 1087 problem solving and tool use of kea and New Caledonian crows in a multi access box paradigm. *PLoS*  
 1088 *ONE*, *6*(6), e20231. <https://doi.org/10.1371/journal.pone.0020231>
- 1089 Bartoń, K. (2020). *MuMIn: Multi-model inference*. <https://CRAN.R-project.org/package=MuMIn>
- 1090 Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4.  
 1091 *Journal of Statistical Software*, *67*(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- 1092 Bates, D., Maechler, M., & Bolker, B. (2012). *lme4: Linear mixed-effects models using S4 classes (2011)*. *R*  
 1093 *package version 0.999375-42*.
- 1094 Bergstrom, C. T., & Lachmann, M. (2004). Shannon information and biological fitness. *Information Theory*  
 1095 *Workshop, 2004. IEEE*, 50–54.
- 1096 Blaisdell, A. P., & Cook, R. G. (2005). Integration of spatial maps in pigeons. *Animal Cognition*, *8*(1), 7–16.
- 1097 Bussey, T. J., Padain, T. L., Skillings, E. A., Winters, B. D., Morton, A. J., & Saksida, L. M. (2008).  
 1098 The touchscreen cognitive testing method for rodents: How to get the best out of your rat. *Learning &*  
 1099 *Memory*, *15*(7), 516–523.
- 1100 Chow, P. K. Y., Lea, S. E., & Leaver, L. A. (2016). How practice makes perfect: The role of persistence,  
 1101 flexibility and learning in problem-solving efficiency. *Animal Behaviour*, *112*, 273–283. [https://doi.org/](https://doi.org/10.1016/j.anbehav.2015.11.014)  
 1102 [10.1016/j.anbehav.2015.11.014](https://doi.org/10.1016/j.anbehav.2015.11.014)
- 1103 Cook, R. G., Geller, A. I., Zhang, G.-R., & Gowda, R. (2004). Touchscreen-enhanced visual learning in rats.  
 1104 *Behavior Research Methods, Instruments, & Computers*, *36*(1), 101–106.
- 1105 Diquelou, M. C., Griffin, A. S., & Sol, D. (2015). *The role of motor diversity in foraging innovations: A*  
 1106 *cross-species comparison in urban birds*.
- 1107 Drayton, L. A., & Santos, L. R. (2014). Insights into intraspecies variation in primate prosocial behavior:  
 1108 Capuchins (*cebus apella*) fail to show prosociality on a touchscreen task. *Behavioral Sciences*, *4*(2),  
 1109 87–101.
- 1110 Eddelbuettel, D., & François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical*  
 1111 *Software*, *40*(8), 1–18. <https://doi.org/10.18637/jss.v040.i08>
- 1112 Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using g\* power  
 1113 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. [https:](https://doi.org/10.3758/BRM.41.4.1149)  
 1114 [//doi.org/10.3758/BRM.41.4.1149](https://doi.org/10.3758/BRM.41.4.1149)
- 1115 Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G\* power 3: A flexible statistical power analysis  
 1116 program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191.  
 1117 <https://doi.org/10.3758/BF03193146>
- 1118 Federspiel, I. G., Garland, A., Guez, D., Bugnyar, T., Healy, S. D., Güntürkün, O., & Griffin, A. S. (2017).  
 1119 Adjusting foraging strategies: A comparison of rural and urban common mynas (*acridotheres tristis*).  
 1120 *Animal Cognition*, *20*(1), 65–74.
- 1121 Gabry, J., & Češnovar, R. (2021). *Cmdstanr: R interface to 'CmdStan'*.
- 1122 Gamer, M., Lemon, J., Gamer, M. M., Robinson, A., & Kendall's, W. (2012). Package “irr.” *Various*  
 1123 *Coefficients of Interrater Reliability and Agreement*.
- 1124 Griffin, A. S., & Guez, D. (2014). Innovation and problem solving: A review of common mechanisms.  
 1125 *Behavioural Processes*, *109*, 121–134. <https://doi.org/10.1016/j.beproc.2014.08.027>
- 1126 Griffin, A. S., Guez, D., Federspiel, I., Diquelou, M., & Lermite, F. (2016). Invading new environments:  
 1127 A mechanistic framework linking motor diversity and cognition to establishment success. *Biological*  
 1128 *Invasions and Animal Behaviour*, 26e46.
- 1129 Griffin, A. S., Guez, D., Lermite, F., & Patience, M. (2013). Tracking changing environments: Innovators  
 1130 are fast, but not flexible learners. *PloS One*, *8*(12), e84907.
- 1131 Hadfield, J. (2010). MCMC methods for multi-response generalized linear mixed models: The MCMCglmm  
 1132 r package. *Journal of Statistical Software*, *33*(2), 1–22. <https://doi.org/10.18637/jss.v033.i02>
- 1133 Hadfield, J. (2014). *MCMCglmm course notes*. [http://cran.r-project.org/web/packages/MCMCglmm/](http://cran.r-project.org/web/packages/MCMCglmm/vignettes/CourseNotes.pdf)  
 1134 [vignettes/CourseNotes.pdf](http://cran.r-project.org/web/packages/MCMCglmm/vignettes/CourseNotes.pdf)

- 1135 Hartig, F. (2019). *DHARMA: Residual diagnostics for hierarchical (multi-level / mixed) regression models*.  
1136 <http://florianhartig.github.io/DHARMA/>
- 1137 Hernán, M. A., & Robins, J. M. (2006). Instruments for causal inference: An epidemiologist's dream?  
1138 *Epidemiology*, 360–372.
- 1139 Homberg, J. R., Pattij, T., Janssen, M. C., Ronken, E., De Boer, S. F., Schoffelmeer, A. N., & Cuppen, E.  
1140 (2007). Serotonin transporter deficiency in rats improves inhibitory control but not behavioural flexibility.  
1141 *European Journal of Neuroscience*, 26(7), 2066–2073.
- 1142 Hutcheon, J. A., Chiolero, A., & Hanley, J. A. (2010). Random measurement error and regression dilution  
1143 bias. *Bmj*, 340, c2289. <https://doi.org/10.1136/bmj.c2289>
- 1144 Isden, J., Panayi, C., Dingle, C., & Madden, J. (2013). Performance in cognitive and problem-solving tasks  
1145 in male spotted bowerbirds does not correlate with mating success. *Animal Behaviour*, 86(4), 829–838.
- 1146 Kangas, B. D., & Bergman, J. (2017). Touchscreen technology in the study of cognition-related behavior.  
1147 *Behavioural Pharmacology*, 28(8), 623. <https://doi.org/10.1097/FBP.0000000000000356>
- 1148 Lajeunesse, M. J., Koricheva, J., Gurevitch, J., & Mengersen, K. (2013). Recovering missing or partial data  
1149 from studies: A survey of conversions and imputations for meta-analysis. *Handbook of Meta-Analysis in*  
1150 *Ecology and Evolution*, 195–206.
- 1151 Lea, S. E., Chow, P. K., Leaver, L. A., & McLaren, I. P. (2020). Behavioral flexibility: A review, a model,  
1152 and some exploratory tests. *Learning & Behavior*, 48(1), 173–187.
- 1153 Lefebvre, L., Whittle, P., Lascaris, E., & Finkelstein, A. (1997). Feeding innovations and forebrain size in  
1154 birds. *Animal Behaviour*, 53(3), 549–560. <https://doi.org/10.1006/anbe.1996.0330>
- 1155 Lin, G. (2020). *Reactable: Interactive data tables based on 'react table'*. [https://CRAN.R-project.org/  
1156 package=reactable](https://CRAN.R-project.org/package=reactable)
- 1157 Liu, Y., Day, L. B., Summers, K., & Burmeister, S. S. (2016). Learning to learn: Advanced behavioural  
1158 flexibility in a poison frog. *Animal Behaviour*, 111, 167–172.
- 1159 Logan, C. J. (2016). Behavioral flexibility in an invasive bird is independent of other behaviors. *PeerJ*, 4,  
1160 e2215.
- 1161 Logan, C. J., Avin, S., Boogert, N., Buskell, A., Cross, F. R., Currie, A., Jelbert, S., Lukas, D., Mares, R.,  
1162 Navarrete, A. F., et al. (2018). Beyond brain size: Uncovering the neural correlates of behavioral and  
1163 cognitive specialization. *Comparative Cognition & Behavior Reviews*.
- 1164 Logan, C. J., McCune, K., MacPherson, M., Johnson-Ulrich, Z., Rowney, C., Seitz, B., Blaisdell, A., Deffner,  
1165 D., & Wascher, C. (2021). *Are the more flexible great-tailed grackles also better at behavioral inhibition?*  
1166 <https://doi.org/10.31234/osf.io/vpc39>
- 1167 Logan, C. J., Shaw, R., Lukas, D., & McCune, K. B. (2022). *How to succeed in human modified environments*.  
1168 <http://corinalogan.com/ManyIndividuals/mil.html>
- 1169 Logan, CJ, McCune, KB, Chen, N, & Lukas, D. (2020). Implementing a rapid geographic range expansion  
1170 - the role of behavior and habitat changes. *In Principle Acceptance by PCI Ecology of the Version on 6*  
1171 *Oct 2020*. <http://corinalogan.com/Preregistrations/gxpobbehaviorhabitat.html>
- 1172 Logan, C., Lukas, D., Blaisdell, A., Johnson-Ulrich, Z., MacPherson, M., Seitz, B., Sevchik, A., & McCune,  
1173 K. (2023). Data: Is behavioral flexibility manipulatable and, if so, does it improve flexibility and problem  
1174 solving in a new context? *Knowledge Network for Biocomplexity, Data package*. [https://doi.org/10.5063/  
1175 FIBR8QNC](https://doi.org/10.5063/FIBR8QNC)
- 1176 Lukas, D., McCune, K., Blaisdell, A., Johnson-Ulrich, Z., MacPherson, M., Seitz, B., Sevchik, A., & Logan,  
1177 C. (2022). Behavioral flexibility is manipulatable and it improves flexibility and problem solving in a new  
1178 context: Post-hoc analyses of the components of behavioral flexibility. *EcoEvoRxiv*. [https://doi.org/10.  
1179 32942/osf.io/4ycps](https://doi.org/10.32942/osf.io/4ycps)
- 1180 Manrique, H. M., Völter, C. J., & Call, J. (2013). Repeated innovation in great apes. *Animal Behaviour*,  
1181 85(1), 195–202. <https://doi.org/10.1016/j.anbehav.2012.10.026>
- 1182 McCune, K., Blaisdell, A., Johnson-Ulrich, Z., Lukas, D., MacPherson, M., Seitz, B., Sevchik, A., & Logan,  
1183 C. (2022). Repeatability of performance within and across contexts measuring behavioral flexibility.  
1184 *EcoEvoRxiv*. <https://doi.org/10.32942/osf.io/kevqp>
- 1185 McCune, KB, MacPherson, M, Rowney, C, Bergeron, L, Folsom, M, & Logan, C. (2019). Is behavioral flexi-  
1186 bility linked with exploration, but not boldness, persistence, or motor diversity? *In Principle Acceptance*  
1187 *by PCI Ecology of the Version on 27 Mar 2019*. [http://corinalogan.com/Preregistrations/g\\_Exploration.  
1188 html](http://corinalogan.com/Preregistrations/g_exploration.html)

1189 McElreath, R. (2016). *Statistical rethinking: A bayesian course with examples in r and stan*. CRC Press.  
1190 <https://doi.org/10.1201/9781315372495>

1191 McElreath, R. (2018). *Statistical rethinking: A bayesian course with examples in r and stan*. Chapman;  
1192 Hall/CRC.

1193 McElreath, R. (2020). *Rethinking: Statistical rethinking book package*.

1194 McInerney, R. E. (2010). Multi-armed bandit bayesian decision making. *Univ. Oxford, Oxford, Tech. Rep*.

1195 Mikhalevich, I., Powell, R., & Logan, C. (2017). Is behavioural flexibility evidence of cognitive complexity?  
1196 How evolution can inform comparative cognition. *Interface Focus*, 7(3), 20160121. <https://doi.org/10.1098/rsfs.2016.0121>

1197

1198 O’Hara, M., Huber, L., & Gajdon, G. K. (2015). The advantage of objects over images in discrimination  
1199 and reversal learning by kea, nestor notabilis. *Animal Behaviour*, 101, 51–60.

1200 R Core Team. (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical  
1201 Computing. <https://www.R-project.org>

1202 Rayburn-Reeves, R. M., Stagner, J. P., Kirk, C. R., & Zentall, T. R. (2013). Reversal learning in rats  
1203 (*rattus norvegicus*) and pigeons (*columba livia*): Qualitative differences in behavioral flexibility. *Journal*  
1204 *of Comparative Psychology*, 127(2), 202.

1205 Sawa, K., Leising, K. J., & Blaisdell, A. P. (2005). Sensory preconditioning in spatial learning using a touch  
1206 screen task in pigeons. *Journal of Experimental Psychology: Animal Behavior Processes*, 31(3), 368.

1207 Schusterman, R. J. (1962). Transfer effects of successive discrimination-reversal training in chimpanzees.  
1208 *Science*, 137(3528), 422–423.

1209 Seitz, B. M., McCune, K., MacPherson, M., Bergeron, L., Blaisdell, A. P., & Logan, C. J. (2021). Using  
1210 touchscreen equipped operant chambers to study animal cognition. Benefits, limitations, and advice.  
1211 *PLoS One*, 16(2), e0246446.

1212 Shaw, R. C., Boogert, N. J., Clayton, N. S., & Burns, K. C. (2015). Wild psychometrics: Evidence for  
1213 ‘general’ cognitive performance in wild new zealand robins, *petroica longipes*. *Animal Behaviour*, 109,  
1214 101–111.

1215 Sol, D., Duncan, R. P., Blackburn, T. M., Cassey, P., & Lefebvre, L. (2005). Big brains, enhanced cognition,  
1216 and response of birds to novel environments. *Proceedings of the National Academy of Sciences of the*  
1217 *United States of America*, 102(15), 5460–5465. <https://doi.org/10.1073/pnas.0408145102>

1218 Sol, D., & Lefebvre, L. (2000). Behavioural flexibility predicts invasion success in birds introduced to new  
1219 zealand. *Oikos*, 90(3), 599–605. <https://doi.org/10.1034/j.1600-0706.2000.900317.x>

1220 Sol, D., Székely, T., Liker, A., & Lefebvre, L. (2007). Big-brained birds survive better in nature. *Proceedings*  
1221 *of the Royal Society of London B: Biological Sciences*, 274(1611), 763–769.

1222 Sol, D., Timmermans, S., & Lefebvre, L. (2002). Behavioural flexibility and invasion success in birds. *Animal*  
1223 *Behaviour*, 63(3), 495–502.

1224 Spence, K. W. (1936). The nature of discrimination learning in animals. *Psychological Review*, 43(5), 427.

1225 Stan Development Team. (2020). *RStan: The R interface to Stan*. <http://mc-stan.org/>

1226 Summers, J., Lukas, D., Logan, C., & Chen, N. (2023). The role of climate change and niche shifts in  
1227 divergent range dynamics of a sister-species pair. *Peer Community Journal*. [https://doi.org/10.24072/](https://doi.org/10.24072/pcjournal.248)  
1228 [pcjournal.248](https://doi.org/10.24072/pcjournal.248)

1229 Ushey, K., Allaire, J., Wickham, H., & Ritchie, G. (2020). *Rstudioapi: Safely access the RStudio API*.  
1230 <https://CRAN.R-project.org/package=rstudioapi>

1231 Warren, J. (1965). Primate learning in comparative perspective. *Behavior of Nonhuman Primates*, 1,  
1232 249–281.

1233 Warren, J. (1966). Reversal learning and the formation of learning sets by cats and rhesus monkeys. *Journal*  
1234 *of Comparative and Physiological Psychology*, 61(3), 421.

1235 Warren, J. M. (1965). The comparative psychology of learning. *Annual Review of Psychology*, 16(1), 95–118.

1236 Wehtje, W. (2003). The range expansion of the great-tailed grackle (*quiscalus mexicanus gmelin*) in north  
1237 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)  
1238 [2003.00970.x](https://doi.org/10.1046/j.1365-2699.2003.00970.x)

1239 Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. [https://](https://ggplot2.tidyverse.org)  
1240 [ggplot2.tidyverse.org](https://ggplot2.tidyverse.org)

1241 Wickham, H., François, R., Henry, L., & Müller, K. (2021). *Dplyr: A grammar of data manipulation*.  
1242 <https://CRAN.R-project.org/package=dplyr>

1243 Wilke, C. (2017). Cowplot: Streamlined plot theme and plot annotations for “ggplot2.” R package version  
1244 0.9. 2; 2017. URL <https://CRAN.R-Project.Org/Package=Cowplot>.

1245 Wolf, J. E., Urbano, C. M., Ruprecht, C. M., & Leising, K. J. (2014). Need to train your rat? There is an  
1246 app for that: A touchscreen behavioral evaluation system. *Behavior Research Methods*, 46(1), 206–214.

1247 Wright, T. F., Eberhard, J. R., Hobson, E. A., Avery, M. L., & Russello, M. A. (2010). Behavioral flexibility  
1248 and species invasions: The adaptive flexibility hypothesis. *Ethology Ecology & Evolution*, 22(4), 393–404.

1249 Xie, Y. (2013). Knitr: A general-purpose package for dynamic report generation in r. *R Package Version*,  
1250 1(7).

1251 Xie, Y. (2017). *Dynamic documents with r and knitr*. Chapman; Hall/CRC.

1252 Xie, Y. (2018). Knitr: A comprehensive tool for reproducible research in r. In *Implementing reproducible*  
1253 *research* (pp. 3–31). Chapman; Hall/CRC.

1254 Xie, Y. (2019). *formatR: Format r code automatically*. <https://CRAN.R-project.org/package=formatR>

1255 Zhu, H. (2021). *kableExtra: Construct complex table with 'kable' and pipe syntax*. [https://CRAN.R-](https://CRAN.R-project.org/package=kableExtra)  
1256 [project.org/package=kableExtra](https://CRAN.R-project.org/package=kableExtra)