

## Aberystwyth University

### *The repeatability of cognitive performance:*

Cauchoix, M.; Chow, P. K. Y.; van Horik, J. O.; Atance, C. M.; Barbeau, E. J.; Barragan-Jason, G.; Bize, P.; Boussard, A.; Buechel, S. D.; Cabirol, A.; Cauchard, L.; Claidière, N.; Dalesman, Sarah; Devaud, J. M.; Didic, M.; Doligez, B.; Fagot, J.; Fichtel, C.; Henke-von der Malsburg, J.; Hermer, E.

#### *Published in:*

Philosophical Transactions of the Royal Society B: Biological Sciences

#### *Publication date:*

2018

#### *Citation for published version (APA):*

Cauchoix, M., Chow, P. K. Y., van Horik, J. O., Atance, C. M., Barbeau, E. J., Barragan-Jason, G., ... Morand-Ferron, J. (2018). The repeatability of cognitive performance: a meta-analysis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1756).

#### **General rights**

Copyright and moral rights for the publications made accessible in the Aberystwyth Research Portal (the Institutional Repository) are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the Aberystwyth Research Portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the Aberystwyth Research Portal

#### **Take down policy**

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

tel: +44 1970 62 2400  
email: [is@aber.ac.uk](mailto:is@aber.ac.uk)

# 1 **The repeatability of cognitive performance: a meta-analysis**

2

3 Cauchoix M<sup>1,2\*</sup>, Chow PKY<sup>3,4\*</sup>, van Horik JO<sup>3\*</sup>, Atance CM<sup>5</sup>, Barbeau EJ<sup>6</sup>, Barragan-Jason  
4 G<sup>2</sup>, Bize P<sup>7</sup>, Boussard A<sup>8</sup>, Buechel SD<sup>8</sup>, Cabirol A<sup>9</sup>, Cauchard L<sup>10</sup>, Claidière N<sup>11</sup>, Dalesman  
5 S<sup>12</sup>, Devaud JM<sup>9</sup>, Didic M<sup>13</sup>, Doligez B<sup>14</sup>, Fagot J<sup>11</sup>, Fichtel C<sup>15</sup>, Henke-von der Malsburg J<sup>15</sup>,  
6 Hermer E<sup>16</sup>, Huber L<sup>17</sup>, Huebner F<sup>15</sup>, Kappeler PM<sup>15,18</sup>, Klein S<sup>9</sup>, Langbein J<sup>19</sup>, Langley EJG<sup>3</sup>,  
7 Lea SEG<sup>3</sup>, Lihoreau M<sup>9</sup>, Lovlie H<sup>20</sup>, Matzel LD<sup>21</sup>, Nakagawa S<sup>22</sup>, Nawroth C<sup>19</sup>, Oesterwind  
8 S<sup>23</sup>, Sauce B<sup>21</sup>, Smith E<sup>24</sup>, Sorato E<sup>20</sup>, Tebbich S<sup>25</sup>, Wallis LJ<sup>17,26</sup>, Whiteside MA<sup>3</sup>, Wilkinson  
9 A<sup>24</sup>, Chaine AS<sup>1,2§</sup>, Morand-Ferron J<sup>16§</sup>.

10

11 <sup>1</sup>Station d'Ecologie Théorique et Expérimentale du CNRS UMR5321, Evolutionary Ecology Group, 2 route du  
12 CNRS, 09200, Moulis, France.

13 <sup>2</sup>Institute for Advanced Studies in Toulouse, 21 allée de Brienne, 31015, Toulouse, France

14 <sup>3</sup>Centre for Research in Animal Behaviour, Psychology, University of Exeter, U.K.

15 <sup>4</sup>Graduate School of Environmental Science, Division of Biosphere Science, Hokkaido University, Sapporo,  
16 Hokkaido, Japan

17 <sup>5</sup>School of Psychology, University of Ottawa, Ottawa, Canada

18 <sup>6</sup>Centre de recherche Cerveau et Cognition, UPS-UMR5549, Toulouse, France

19 <sup>7</sup>Institute of Biological and Environmental Sciences, University of Aberdeen, Aberdeen, U.K.

20 <sup>8</sup>Department of Zoology/Ethology, Stockholm University, Svante Arrheniusväg 18B, 10691 Stockholm, Sweden

21 <sup>9</sup>Research Center on Animal Cognition (CRCA), Center for Integrative Biology (CBI); CNRS, University Paul  
22 Sabatier, Toulouse, France

23 <sup>10</sup>Département de Sciences Biologiques, Université de Montréal, Montreal, QC, Canada

24 <sup>11</sup>Aix Marseille University, CNRS, LPC, Marseille, France

25 <sup>12</sup>Institute of Biological, Environmental and Rural Sciences, Aberystwyth University, Aberystwyth, U.K.

26 <sup>13</sup>AP-HM Timone & Institut de Neurosciences des Systèmes, Marseille, France

27 <sup>14</sup>CNRS UMR 5558, Université Lyon 1, Department of Biometry and Evolutionary Biology, France

28 <sup>15</sup>Behavioral Ecology & Sociobiology Unit, German Primate Center, Göttingen, Germany

29 <sup>16</sup>Department of Biology, University of Ottawa, Ottawa, Canada

30 <sup>17</sup>Comparative Cognition, Messerli Research Institute, University of Veterinary Medicine Vienna, Medical  
31 University of Vienna, University of Vienna, Vienna, Austria

32 <sup>18</sup>Department of Sociobiology/ Anthropology, University of Göttingen, Göttingen, Germany

33 <sup>19</sup>Institute of Behavioural Physiology, Leibniz Institute for Farm Animal Biology, Dummerstorf, Germany

34 <sup>20</sup>IFM Biology, Linköping University, 58183 Linköping, Sweden

35 <sup>21</sup>Department of Psychology, Rutgers University, Piscataway, USA

36 <sup>22</sup>Evolution & Ecology Research Centre, School of Biological, Earth & Environmental Sciences, University of  
37 New South Wales, Sydney, NSW 2052, Australia

38 <sup>23</sup>Faculty of Agricultural and Environmental Sciences, University of Rostock, Rostock, Germany

39 <sup>24</sup>School of Life Sciences, University of Lincoln, Lincoln, U.K.

40 <sup>25</sup>Department of Behavioural Biology, University of Vienna, Austria

41 <sup>26</sup>Department of Ethology, Eötvös Loránd University, Budapest, Hungary

42

43 \*Shared first authorship listed alphabetically

44 §Shared senior authorship listed alphabetically

45

46 Corresponding author: Maxime Cauchoix (mcauchoixxx@gmail.com)

47 Author Contributions: MC, PKYC, JOvH, ASC, SEGL, and JM-F defined research; all  
48 authors except SN contributed primary data either for the initial or final manuscript, MC  
49 conducted analyses and SN provided code and commented on analyses; MC, PKYC, and  
50 JOvH wrote the manuscript with contributions from ASC and JM-F. Authors who contributed  
51 data wrote their respective methods sections for the supporting information. All authors read  
52 and commented on the manuscript.

53

54

55

56

57

58

59

60

61

62

63

64

65

66

67

68

69

70

71

72

73

74

75

76

77

78

79

80

81 **ABSTRACT**

82 Behavioural and cognitive processes play important roles in mediating an individual's  
83 interactions with its environment. Yet, while there is a vast literature on repeatable individual  
84 differences in behaviour, relatively little is known about the repeatability of cognitive  
85 performance. To further our understanding of the evolution of cognition, we gathered 44  
86 studies on individual performance of 25 species across six animal classes and used meta-  
87 analysis to assess whether cognitive performance is repeatable. We compared repeatability  
88 (R) in performance (1) on the same task presented at different times (temporal repeatability),  
89 and (2) on different tasks that measured the same putative cognitive ability (contextual  
90 repeatability). We also addressed whether R estimates were influenced by seven extrinsic  
91 factors (moderators): type of cognitive performance measurement, type of cognitive task,  
92 delay between tests, origin of the subjects, experimental context, taxonomic class and  
93 publication status. We found support for both temporal and contextual repeatability of  
94 cognitive performance, with mean R estimates ranging between 0.15 and 0.28. Repeatability  
95 estimates were mostly influenced by the type of cognitive performance measures and  
96 publication status. Our findings highlight the widespread occurrence of consistent inter-  
97 individual variation in cognition across a range of taxa which, like behaviour, may be  
98 associated with fitness outcomes.

99

100 *Keywords:* cognitive repeatability; evolutionary biology of cognition; individual differences;  
101 learning; memory; attention.

## 102 INTRODUCTION

103

104 Cognition has been broadly defined as the acquisition, processing, storage and use of  
105 information [1], and hence plays an important role in mediating how animals behave and  
106 interact with their environment. While comparative studies have broadened our understanding  
107 of how socio-ecological selection pressures shape cognitive evolution [2–4], relatively little is  
108 known about the adaptive significance of inter-individual variation of cognitive abilities [5,6].  
109 There is however some evidence that learning may be under selection if it influences fitness  
110 [6-19]. Opportunities to learn have been linked to increased growth rate [7], and individual  
111 learning speed can correlate with foraging success [8,9]. Greater cognitive capacities may  
112 allow individuals to better detect and evade predators [10,11] and may also influence their  
113 reproductive success [12–15]; but see [16]. Finally, rapid evolutionary change in learning  
114 abilities have also been shown by experimentally manipulating environmental conditions,  
115 revealing trade-offs between fitness benefits and costs to learning [17–20]. Accordingly, we  
116 might expect selection to act on individual differences in cognitive ability in other species and  
117 contexts.

118

119 As selection acts on variation, a fundamental prerequisite to understanding the evolution of  
120 cognition in extant populations requires an assessment of individual variation in cognitive  
121 traits [21]. The approach most commonly used in evolutionary and ecological studies to  
122 estimate consistent among-individual variation has its origin in quantitative genetics [22,23].  
123 This approach compares the variation in two or more measures of the same individual, with  
124 variation in the same trait across all individuals to distinguish between variation due to  
125 “noise” and variation among individuals. The amount of variation explained by inter-  
126 individual variation relative to intra-individual variation is termed the “intra-class correlation  
127 coefficient” or “repeatability” (R). Repeatability coefficients are often used to estimate the  
128 upper limit of heritability [23] but see [22], and thus quantifying repeatability is a useful first  
129 step in evolutionary studies of traits [24].

130

131 Assessing the repeatability of behavioural or cognitive traits is, however, challenging, because  
132 the context of measurement can influence the behaviour of animals, and thus, the value  
133 recorded. Contextual variation can come from the internal state of the organism (e.g. hunger,  
134 circadian cycle, recent interactions, stress) and/or the external environment, which may differ  
135 between trials [25]. Moreover, behavioural and cognitive measures may suffer further

136 variation between measures as experience with one type of measure or test can influence  
137 subsequent measures via processes such as learning and memory [26]. While this issue has  
138 been recognised and discussed in recent research on animal personality [27], it may be  
139 particularly relevant when assaying the repeatability of cognitive traits. Consequently, we  
140 might therefore expect higher within-individual variation in behavioural or cognitive  
141 measures compared with morphological or physiological measures, due to greater differences  
142 in the context (internal and/or external) of repeated sampling.

143  
144 Research on animal personality has provided a broad understanding that individual  
145 differences in behaviour are repeatable (average  $R = 0.37$ ) across time and contexts [28],  
146 hence revealing an important platform for selection to act on [29–32]. Yet, relatively little is  
147 known about the stability of inter-individual variation in cognitive traits, such as those  
148 associated with learning and memory [26]. Some examples of repeatability estimates suggest  
149 that children show good test–retest reliability on false-belief tasks used to assess theory-of-  
150 mind [26,33]. Consistent individual differences in performance on cognitive tasks have also  
151 been documented in a few non-human animals, such as guinea pigs, *Cavia aperea f. porcellus*  
152 [34,35], zebra finch, *Taenopigya guttata* [36], Australian magpies, *Gymnorhina tibicen* [37],  
153 mountain chickadees, *Poecile gambeli* [38], bumblebees, *Bombus terrestris* [39] and snails,  
154 *Lymnaea stagnalis* [40]. While the paucity of repeatability measures of cognitive performance  
155 may stem from the recency of interest in the evolutionary ecology of cognitive traits [41,42],  
156 it may also suggest that it is difficult to accurately capture repeatable measures of cognitive  
157 ability [43]. Further investigation into the consistency of individual differences in cognition  
158 and how internal and external factors may influence repeatability estimates of these measures  
159 is therefore warranted.

160  
161 Recent advances in analytical techniques, such as the use of mixed-effect models, have  
162 facilitated the assessment of repeatability of behavioural traits, by accounting for the potential  
163 confounding effects of both internal and external contextual variations [44,45]. Such  
164 approaches can help provide more accurate estimates of repeatability of cognitive traits and  
165 could provide new insights to the influence of internal and external factors on cognitive  
166 performance. For example, we can now explicitly address the effect of time, or an  
167 individual's condition, on the repeatability of traits of interest such as learning performance.  
168 Likewise, we can examine the effect of external factors, for example by modeling the  
169 environment (e.g. group size at testing) or the type of test employed (e.g. spatial vs. colour

170 cues in associative learning). Adopting these methods (i.e. adjusted repeatability [46]) could  
171 therefore facilitate studies that generate repeatability estimates of cognitive performance and  
172 provide greater clarity into the sources of variation in measures of cognition in this rapidly  
173 expanding field.

174

175 In this study, we collated 38 unpublished datasets (see below) and used R values that are  
176 reported in 6 published studies to conduct a meta-analysis. We aim to (1) estimate average  
177 repeatability of cognitive performance across different taxa, and (2) discuss the implications  
178 of how internal and external factors influence measures of cognitive repeatability. To do this,  
179 we first assessed individual performances from 14 different cognitive tasks from 25 species of  
180 six animal classes. For each of the 14 tasks, we assessed multiple performance measures, such  
181 as number of trials to reach a criterion or success-or-failure for the same task. We then  
182 assessed *temporal repeatability* by comparing individual performances on multiple exposures  
183 of the same task, and *contextual repeatability* by comparing individual performances on  
184 different tasks that measure the same putative cognitive ability. We also used meta-analysis to  
185 investigate whether there are general across-taxa patterns of repeatability for different tasks  
186 and which factors (type of cognitive performance measurement, type of cognitive task, delay  
187 between tasks, origin of the subjects, experimental context, taxonomic class, and whether the  
188 R value was published or unpublished) might influence the repeatability of cognitive  
189 performance.

190

## 191 **METHODS**

### 192 **Data collection**

193 We followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses  
194 (PRISMA) approach for the collation of the datasets used in the current study [47]. We first  
195 collected published repeatability estimates of cognitive performance (Figure S1). We did not  
196 include studies reporting inter-class correlations (Pearson or Spearman) between cognitive  
197 performances on tasks measuring different cognitive abilities (i.e., general intelligence or ‘g’)  
198 as we considered these outside the scope of this meta-analysis. Although we acknowledge that  
199 results from the literature on test-retest [48,49] or convergent validity [50] in psychology  
200 would be relevant to compare with the present study, we also considered them beyond the  
201 scope of this paper as their inclusion would have led to a heavy bias towards studies on  
202 humans. We only found 6 publications reporting repeatability values for cognitive  
203 performance (R) in 6 different species: 1 arachnid [51], 2 mammals [52–54] and 3 birds

204 [15,55,56], with a sample size ranging from 15 to 347 (Mean: 54.7, Median: 33) and number  
205 of repeated tests varying from 2 to 4 (Mean: 2.5, Median: 2).

206  
207 To complement our dataset from published studies, we used an ‘individual-patient-data’ meta-  
208 analysis approach commonly used in medical research [57] in which effect sizes are extracted  
209 using the same analysis on primary data [57]. We invited participants from a workshop on the  
210 ‘Causes and consequences of individual variation in cognitive ability’ (36 people), as well as  
211 25 colleagues working on individual differences in cognition, to contribute primary datasets  
212 of repeated measurements of cognitive performance. From this approach, we assembled 38  
213 primary datasets from unpublished (9 datasets: 6 were fully unpublished while 3 had similar  
214 methods published from the same laboratory group) or published sources (29 datasets:  
215 including repeated measures of cognitive performance but that didn’t report R values) that we  
216 could use to compute repeatability using consistent analytical methods (Figure S1, see shared  
217 repository link). These datasets comprised 20 different species of mammals (humans  
218 included), insects, molluscs, reptiles and birds (Table S1 and Table S2). Details about  
219 subjects, experimental context and cognitive tasks for each dataset can be found in electronic  
220 supplementary material (ESM methods).

221  
222 Each dataset included 4 – 375 individuals (Mean: 46.6, Median: 29), that performed 2 – 80  
223 (Mean: 7.9, Median: 2) repetitions of tests targeting the same cognitive process, either by  
224 conducting the same task presented at different points in time (*temporal repeatability*, see  
225 Table S1), or different tasks aimed at assessing the same underlying cognitive process but  
226 using a different protocol (*contextual repeatability*, see Table S2). Tasks considered to assess  
227 contextual repeatability differed by stimulus dimension (e.g. spatial vs. colour reversal  
228 learning in Cauchoix- great tit dataset), sensory modality (e.g. visual vs. olfactory  
229 discrimination in Henke- v.d. Malsburg -microcebus dataset), change in experimental  
230 apparatus (e.g. colour discrimination on touch screen and on solid objects in Chow-squirrel  
231 lab dataset) or could be a different task designed to measure the same cognitive process (i.e.  
232 Mouse Stroop Test and the Dual Radial Arm Maze to measure external attention in Matzel-  
233 attention mice dataset).

234

### 235 **Repeatability analysis for primary data**

236 All analyses were performed in the R environment for statistical computing version 3.3.3  
237 [58]. We performed the same repeatability analysis for all primary data provided by co-



238 authors: (1) We first transformed cognitive variables to meet assumptions of normality; (2) To  
239 assess whether time-related changes (i.e. the number of repetitions of the same task or test  
240 order of different tasks), and/or an individual's sex and age (hereafter, individual  
241 determinants) played a role in repeatability of cognitive performances, we then computed 3  
242 types of repeatability values with a mixed-effects model approach using the appropriate link  
243 function in the 'rptR' package [59]. Specifically, we calculated unadjusted repeatability (R),  
244 repeatability adjusted for test order (Rn), and repeatability adjusted for test order and  
245 individual determinants (Rni) for *temporal* and *contextual* repeatability separately; (3) For  
246 cases with unadjusted R close to 0 ( $< 0.005$ ), we computed the R estimate using a least  
247 squares ANOVA approach as advised in [60–62] using the 'ICC' package [63]; and (4) we  
248 removed R estimates from further analyses when residuals were not normal or overdispersed  
249 (for Poisson distribution) and for data that could not be transformed to achieve normality. See  
250 ESM general methods for more details.

251

### 252 **Meta-analysis and meta-regression**

253 We collated the 178 R values computed from primary data with the 35 R values from  
254 published studies to obtain a total of 213 estimates of cognitive repeatability. We did not  
255 recompute repeatability de novo for published studies that provide repeatability values as the  
256 statistics used in these papers are the same or similar to those used here for primary data (e.g.  
257 mixed-model approach with or without 'rptR' package). We then used a meta-analytic  
258 approach to examine average R estimates across species of cognitive performance. This  
259 approach allowed us to: (1) take into account sample size and number of repeated measure  
260 associated with each R value in the estimation of average cognitive repeatability, (2) control  
261 for repeated samples (i.e., avoid pseudoreplication) of the same species (taxonomic bias), the  
262 same laboratory group (i.e., same senior author; observer bias) or the same experiment  
263 (measurement bias) by including these factors as random effects, and (3) ask whether other  
264 specific factors (fixed effects called "moderators" in meta-analysis, see below) could explain  
265 the variation in repeatability of cognitive tests.

266

267 For each of the 6 types of R analysis (i.e., unadjusted temporal R, adjusted temporal R for test  
268 order, adjusted temporal R for test order and individual determinants, unadjusted contextual  
269 R, adjusted contextual R for test order, adjusted contextual R for test order and individual  
270 determinants), we performed 3 different multilevel meta-analyses by fitting Linear Mixed  
271 Models (LMMs) using the 'metafor' package [64]: (1) a standard meta-analytic model

272 (intercept-only model) to estimate the overall mean effect size, (2) 7 univariate (multilevel)  
273 meta-regression models to independently test the significance of each moderator. For each  
274 model, we used standardized (Fisher's  $Z$  transformed)  $R$  values as the response variable.  
275 Finally, we conducted (3) a type of Egger's regression to test for selection bias.

276

277 In the intercept only model, overall effects (intercepts) were considered statistically  
278 significant if their 95% CIs did not overlap with zero. To examine whether the overall effect  
279 sizes of the 6 different analyses were statistically different from each other, we manually  
280 performed multiple pairwise  $t$ -tests by comparing  $t$  values calculated from meta-analytic  
281 estimates and their standard errors.

282

283 In meta-regression models, we accounted for variance in repeatability of cognitive  
284 performance by adding both fixed and random effects. We accounted for variation in  
285 repeatability related to fixed effects by including moderators. We considered 7 moderators  
286 (detailed in ESM general methods and Figure 1 and 2 captions): type of cognitive  
287 performance measurement (e.g. success or failure, latency, the number of trials before  
288 reaching a learning criterion); type of cognitive task (e.g. reversal learning, discrimination  
289 learning); median delay between tests; experimental context (conducted in the wild or in  
290 captivity); the origin of subjects (wild or hand raised), taxonomic class, and publication status  
291 (whether the  $R$  value was published or unpublished). We also took into account non-  
292 independence of data by including random effects, including species (multiple datasets from  
293 the same species), laboratory group (experiments conducted by the same PI), and experiment  
294 (experiments on the same subjects; see ESM general methods for more details).

295

296 We controlled for the possibility that phylogenetic history influences the repeatability of  
297 cognitive abilities (i.e. closely related species may be more likely to show similar estimates of  
298 cognitive repeatability) by using a covariance matrix based on an order-level phylogenetic  
299 tree (using Open Tree of Life [65] and "rotl" R package [66] ) but only in the intercept only  
300 model as meta-regression models failed to converge with this additional information. We ran  
301 the intercept only meta-analysis with and without controlling for the effect of phylogeny and  
302 found that phylogenetic relationships had negligible effects on average repeatability of  
303 cognitive abilities (Table S5), justifying its exclusion in subsequent meta-regression models.

304

305 For meta-regressions, we report conditional  $R^2$  (sensu [67]) which quantifies the proportion  
306 of variance explained by fixed (moderators) and random effects along with p-values from  
307 omnibus tests [64] which test the significance of multiple moderator effects. When omnibus  
308 tests were significant ( $p < 0.05$ ) we ran the same meta-regression model without the intercept  
309 to compute and plot beta coefficients associated with each level of the moderator (Figure S10  
310 and S11) and performed multiple pairwise comparisons to estimate statistical differences  
311 between all combinations of moderator levels. We corrected for multiple comparisons using a  
312 false discovery rate adjustment of p-values [68].

313  
314 We assessed the extent of variation among effect sizes in each meta-analytic model (intercept  
315 only) by calculating heterogeneities ( $I^2$ ). Along with the overall heterogeneity ( $I^2_{\text{total}}$ ), which  
316 represents between-study variance divided by the total variance [69], we also provide  
317 estimates of heterogeneity for each random factor (species, laboratory and experiment)  
318 following [70].  $I^2$  values of 25%, 50% and 75% are generally considered to be low, moderate  
319 and high levels of heterogeneity, respectively [69].

320  
321 Finally, we statistically tested for selection bias in the dataset by conducting a type of Egger's  
322 regression [71]. Given that effect sizes were not always independent from each other (i.e.  
323 some came from the same study), we employed a mixed-model version of Egger's regression  
324 using the full models (7 moderators as fixed effects) with the sampling standard errors (SE) of  
325 each effect size as a moderator [72,73]; a regression slope of the SE significantly different  
326 from zero indicates selection bias [71]. Such a significant effect usually indicates that large  
327 effect sizes with large sampling variance (small sample size) are more prevalent than  
328 expected, potentially overestimating the overall effect size (i.e.,  $R$ ).

329

## 330 **RESULTS**

### 331 *Dataset summary*

332 Repeatability estimates computed from primary data are presented together with published  $R$   
333 values in Table S1 for temporal repeatability and Table S2 for contextual repeatability. For  
334 temporal repeatability, we used 22 studies on 15 species in which 4 to 375 (Mean: 56.3,  
335 Median: 40) individuals performed a median of 2, 95%CI [1.91, 2.11] repeated tests, leading  
336 to a total of 106 repeatability analyses (40  $R$ ; 40  $R_n$ ; and 26  $R_{ni}$ ). For contextual repeatability,  
337 we used 27 studies on 20 species in which 4 to 297 (Mean: 41, Median: 24) individuals

338 performed a median of 2, 95%CI [1.80, 2.15] repeated tests, leading to a total of 107  
339 repeatability analyses (38 R; 32 Rn; and 37 Rni).

340

#### 341 ***Repeatabilities for individual studies***

342 Repeatability of cognitive performance varied widely between studies and was distributed  
343 from negative (i.e. higher within-individual than between-individual variability, computed for  
344 unadjusted R only) to highly positive repeatability (close to 1) for unadjusted R (Figure 1-2  
345 and Figure S2). Confidence intervals also varied greatly among species and cognitive tasks,  
346 particularly for unadjusted R of temporal repeatability (Figure 1) and contextual repeatability  
347 (Figure 2). Such heterogeneity in R between datasets, wide confidence intervals, as well as  
348 high variation in sample size and number of repetitions, suggest that mean estimates would be  
349 better assessed through meta-analysis regression.

350

#### 351 ***Meta-analysis: overall repeatability estimates, heterogeneities and publication bias***

352 We first used meta-analysis (intercept-only) models to compute mean estimates of cognitive  
353 repeatability while accounting for variation in sample size and repetition number between  
354 studies. Intercept-only models revealed significant low to moderate [0.15 - 0.28] mean  
355 estimates of cognitive repeatability across analyses (Table 1, Figure 3). Performing the same  
356 analysis with or without controlling for phylogenetic history suggests that class-level  
357 phylogenetic relationships had little influence on mean cognitive repeatability estimates  
358 (Table S4).

359

360 While confidence intervals of mean repeatability estimates (Figure 3 and Table 1) indicate  
361 considerable variability in the repeatability of cognitive performance between studies,  
362 inconsistency between effect sizes is better captured by heterogeneity  $I^2$  for meta-analysis  
363 [74]. We found moderate to high total heterogeneity ( $32\% < I^2 < 88\%$ , Table 1) as in other  
364 across species meta-analyses [74]. Indeed, a considerable proportion of the total heterogeneity  
365 ( $I^2$  total) is due to variations between species ( $I^2$  species). Using repeatability from different  
366 cognitive measurements in the same experiment ( $I^2$  experiment) also produced a moderate  
367 level of heterogeneity, suggesting that the type of cognitive measurement plays a role in  
368 repeatability estimation.

369

370 We investigated whether our meta-analysis model showed any bias in publication or selection  
371 using a type of Egger's regression. Egger's regressions suggest significant bias for unadjusted

372 temporal R. Such bias is probably related to the high number of low sample size studies. To  
373 further evaluate the robustness of our mean estimates, we ran a sensitivity analysis using a  
374 “leave one out procedure” (ESM general methods) in which we computed mean estimates by  
375 removing a single R value for each R value in the dataset and generating a distribution of  
376 mean estimates. The distribution of “leave one out” mean estimates were concentrated around  
377 the original mean estimate, which suggests that meta-analytic results are not driven by one  
378 particular R value (Figure S10). Finally, we assessed whether mean estimates obtained for  
379 each type of R analysis was significantly different from each other using multiple t-test  
380 comparisons. We found that adjusted temporal R for test order was significantly lower than  
381 other types of R analyses before correcting for multiple comparisons (Table S5). However,  
382 we found no significant differences after correcting for multiple comparisons for all  
383 combinations of R analyses.

384

#### 385 *Meta-regression: effects of moderators*

386 To better understand the factors that influence heterogeneity of repeatability, we included the  
387 type of cognitive performance measurement, the type of cognitive task, median delay between  
388 repetitions, origin of the subjects, experimental context, taxonomic class, and publication  
389 status as moderators in our models of repeatability. Effects of those factors on raw R values  
390 can be inspected visually in Figures S3-9. However, to assess the effects of these factors  
391 while accounting for variation in sample size and repetition number between studies, meta-  
392 analytical tools are necessary. The total number of repeatability values compiled for each type  
393 of R analysis (Table 1) was not sufficient to run a full model to assess the effects of all 7  
394 moderators together. We therefore ran 7 independent univariate (multilevel) meta-regression  
395 models, which revealed that the type of cognitive performance measurement significantly  
396 influenced all types of R values, except for unadjusted temporal values (Table 2), and  
397 accounted for 14 to 100% of the variance (R<sup>2</sup>c). The investigation of beta coefficients  
398 associated with each type of cognitive measurement (Figure S11) suggests that normalized  
399 index (scores computed specifically for the study e.g. Matzel et al., dataset) and success-or-  
400 failure measures are significantly more repeatable for contextual R<sub>ni</sub> estimates than other  
401 types of R analyses. However, as this pattern is not observed for other types of R analyses,  
402 results should be interpreted with caution. Publication status also significantly influenced  
403 contextual repeatability and accounted for 24 to 70% of the variance (Table 2), with published  
404 R values being significantly higher than the R values that are computed from primary data  
405 (Figure S12).

406  
407 We found that the type of cognitive task, median delay between tasks, experimental context,  
408 the origin of the subjects or taxonomic class did not show consistently significant effects  
409 across different types of R analyses. The significant effect of cognitive task type on  
410 unadjusted contextual R should be interpreted cautiously as it is present only for one type of R  
411 analysis and is thus probably not robust (Table 1 and Figure 1). The same is also true for the  
412 marginally significant effect of median delay between tasks; its positive beta coefficient (0.06,  
413 see also Figure S3) suggests that repeatability increased with the delay between tests. This  
414 finding could be driven by high R values from the study by Barbeau et al., in humans (Table  
415 S1) despite a very long median delay between trials (540 days). Indeed, the p-value associated  
416 to median delay became non-significant when running the same meta-regression without  
417 those data.

418

## 419 **DISCUSSION**

420 We aimed to explore the repeatability of cognitive performance across six animal classes. We  
421 examined repeatability by assessing whether inter-individual variation in cognitive  
422 performance was consistent on the same task across two or more points in time (i.e., temporal  
423 repeatability) or whether performances were consistent across different tasks that are designed  
424 to capture the same cognitive process (i.e., contextual repeatability). Overall, our meta-  
425 analysis revealed robust and significant low to moderate repeatability of cognitive  
426 performance ( $R = [0.15 - 0.28]$ ). We found that the type of cognitive performance  
427 measurement (e.g. the number of trials to reach a criterion, latency) affected most estimates of  
428 repeatabilities while the type of cognitive task (e.g. reversal learning, discrimination learning,  
429 mechanical problem solving), delay between task repetitions, the origin of animals  
430 (wild/wild-caught or laboratory-raised/hand-raised), experimental context (in the wild or  
431 laboratory), taxonomic class, and origin of R values (published vs. primary data) did not  
432 consistently show significant effects on R estimates.

433

### 434 *Are measures of cognition repeatable?*

435

436 High plasticity of cognitive processes may result in low or null estimates of repeatability. Yet,  
437 we found a significant, but low, average R estimate for unadjusted temporal repeatability of  
438 cognitive performance ( $R = 0.15$ ). Our highest temporal repeatability estimate adjusted for  
439 test order and individual determinants reached  $R = 0.28$ . Although this estimate remains lower

440 than that observed for animal personality ( $R = 0.37$ ) [75], our findings suggest that individual  
441 variation in performance on the same cognitive task is moderately consistent across time in a  
442 wide range of taxa. This result is particularly striking because internal and external influences  
443 on task performance are unlikely to be identical between trials; such influences should inflate  
444 intra-individual variation between trials, and therefore reduce  $R$ . The results we obtained are  
445 in line with low to moderate heritability estimates of cognitive performance collected on  
446 laboratory populations (reviewed in [76], also see Sauce et al., and Sorato et al., in this issue),  
447 and with selectively bred animals that have shown large differences in, for example,  
448 numerical learning in guppies [77], oviposition learning in *Drosophila* [78] and butterflies  
449 [79], or maze navigation in rats [80]. These findings may promote future investigation of  
450 individual variation in cognitive performance, ideally as a first step towards assessing  
451 heritability, the effect of developmental environment and experience on this variation, and  
452 examining potential evolutionary consequences of this variation [6,81].

453

454 Contextual repeatability was assessed by examining performance on novel variants of the  
455 same task (e.g. change of stimuli dimension) or different tasks that we considered assessed the  
456 same putative cognitive process. The use of different task variants has been advocated to  
457 further improve our understanding of cognitive processes, for instance in the context of  
458 assessing convergent validity of tasks [48, Volter et al., in this issue]. Accordingly, our  
459 estimates of contextual repeatability were moderate ( $R = [0.20-0.27]$ ) and significant,  
460 indicating that the use of different stimuli dimensions, perceptual dimensions, apparatuses and  
461 tests allow accurate measures of repeatable variation of individual cognitive performance.  
462 However, our interpretation of  $R$  values assumes that performance on each cognitive test is  
463 independent of other traits that could be repeatable as well, such as motor capacities,  
464 motivation or personality traits [48].

465

466 Accurate estimates of contextual repeatability may be confounded in tasks that use different  
467 stimuli or perceptual dimensions. For instance, adaptive specialisations that result in  
468 differential attention to particular stimuli may result in high within-individual variation in  
469 performance over contexts, or in low between-individual variation in one or both contexts  
470 [82] (e.g. individuals of some species may show greater variation in their performance when  
471 learning a shape discrimination, but show relatively little variation when learning a colour  
472 discrimination, even if both tasks require visual-cue learning e.g. [83,84]). Using different  
473 tasks or apparatuses to examine the same putative cognitive process may also lead to low

474 contextual repeatability if the salience of stimuli differs between apparatuses. For example,  
475 presenting stimuli on a touchscreen as opposed to presenting stimuli with solid objects may  
476 vary the salience of stimuli [85]. Such differences may inflate within-individual variance and  
477 thus decrease repeatability. Finally, while we may assume similar cognitive processes are  
478 involved in variants of the same task, we may obtain low contextual repeatability if the  
479 variants require different cognitive processes. One possible solution is to conduct repeatability  
480 analyses on the portion of variance likely due to a shared cognitive process by incorporating  
481 measures of ‘micro-behaviours’. For example, Chow and colleagues [86] used the response  
482 latencies to correct and incorrect stimuli to reflect inhibitory control, and the rate of head-  
483 switching (head-turning between stimuli) to reflect attention, alongside using the number of  
484 errors in learning a colour discrimination-reversal learning task on a touch screen. Assessing  
485 micro-behaviours may therefore capture specific processes that are closely related to the  
486 general cognitive process than more classical approaches. Accordingly, assays of repeatability  
487 of cognitive performances could then be examined by repeatedly recording a suite of micro-  
488 behavioural traits as well as traditional measures of performance in the same, or variants of  
489 the same, task.

490

#### 491 ***Test order and the repeatability of cognitive performance***

492 Animals may improve their performance with increased learning/experience of the same task  
493 or on different but related tasks. Hence, controlling for time-related changes (i.e. the number  
494 of repetitions of the same task) or task presentation order (i.e. test order) may produce more  
495 accurate estimates of repeatability [87]. However, while our adjusted estimates of temporal  
496 and contextual repeatability remained significant when controlling for test order, they did not  
497 increase (Table 1, Figure 3). These findings suggest that repetition number, or task order, may  
498 have a negligible influence on repeatability, at least within the range of values represented in  
499 our sample.

500

501 Estimates of temporal repeatability (Table S1) suggest that there may however be an optimal  
502 number of repetitions when estimating individual variation in cognitive performance. Indeed,  
503 prolonged exposure to the same task may reduce most, if not all, between-individual variation  
504 in performance (i.e. individuals reach a plateau in performance with increased experience of  
505 the same task): high repetitions of the same task (ranging from 7 to 80 repetitions) produced  
506 moderate-low repeatability (mean  $R = 0.22$ ) whereas analyses with low repetitions (ranging  
507 from 2 to 3 repetitions) produced a moderate-high repeatability (mean  $R = 0.42$ ).



508 Consequently, increasing the number of measures of cognitive performance strengthens  
509 memory and learning on a given task, which may increase within-individual variance between  
510 tests as internal and external conditions change across repetitions. Likewise, memory and  
511 learning may increase within-individual variance between different tasks due to carry-over  
512 effects. Carry-over effects on repeatability may be controlled by running all tests in the same  
513 order for all subjects, and by including test number or test date for a given task [87]. The  
514 effect of test order on contextual repeatability should however be treated with caution, as it  
515 may be influenced by the number of R estimates based on small sample size studies, and may  
516 also result from GLMM-based repeatability approaches which force R to be positive, in  
517 comparison to unadjusted R. Nevertheless, studying the impact of repetition number or prior  
518 test exposure may help improve our understanding of how experience can influence cognitive  
519 performance.

520

#### 521 *Individual determinants of the repeatability of cognitive performance*

522 The addition of individual effects such as sex and age, when available, appeared to increase  
523 temporal but not contextual repeatability, relative to models that only included test order  
524 (Table 1, Figure 3). This effect on temporal repeatability may partly result from differences in  
525 the processes that underlie performance on cognitive tasks between juveniles and adults. For  
526 example, immature freshwater snails, *Lymnaea stagnalis*, show impaired memory for the  
527 association between a light flash and the whole body withdrawal response until they reach  
528 maturity [88], juvenile Australian magpies, *Cracticus tibicen*, show impaired performance on  
529 a spatial memory task when tested 100 days after fledging than compared to those birds that  
530 were tested 200 and 300 days after fledging [15], and honeybee workers, *Apis mellifera L.*,  
531 show impaired spatial memory when tested under 16 days of age as adults than compared to  
532 their counterparts that were older than 16 days [89]. Adult Eurasian harvest mice, *Micromys*  
533 *minutus*, also show higher repeatability than juveniles on a spatial recognition task [53].  
534 Controlling for age and developmental life-stage, either experimentally (e.g. target one age  
535 group) or statistically, may therefore play an important role in obtaining accurate estimates of  
536 repeatability of cognitive performance.

537

538 Males and females may also experience different selective pressures on given cognitive  
539 processes that reflect different fitness consequences. Examples of such sex differences include  
540 spatial orientation and reference memory in rodents [90], colour and position cues learning in  
541 chicks [91], and foraging innovation in guppies [92]. Sex differences in cognitive processes

542 may result from mating behaviours such as territory defense or mate searching, which may  
543 reduce between-individual variation within the same sex. Here, we have only examined and  
544 discussed a few of the individual factors that can influence measures of cognitive  
545 performance across individuals, and thus potentially impact estimates of repeatability. We  
546 suggest that the choice of variables included in analyses of adjusted repeatability should  
547 reflect the goals of the study, and include explanations of what aspects are controlled for and  
548 more importantly, why [24].

549

### 550 *Moderators of the repeatability of cognitive performance*

551 Variation among studies used in a meta-analysis can cause heterogeneity in effect sizes that  
552 are directly attributable to the experimental approach. Accounting for such variation can  
553 provide insights into which factors influence the trait of interest [74]. For example, we might  
554 expect that repeated measurements that are obtained after shorter time intervals may produce  
555 better estimates of repeatability because the internal and external states of individuals may be  
556 more similar [75]. However, our results suggest that the interval between two tasks had no  
557 influence on most estimates of temporal or contextual repeatability. Although animals may  
558 form memory associations on a given test, our finding suggest a negligible influence of carry-  
559 over effects on the relative extent of between vs. within-individual variation.

560

561 We found that the type of cognitive performance measure had a strong effect on estimates of  
562 repeatability (Table 2). For contextual repeatability, the lowest estimated R values were  
563 obtained for latency measures, with most confidence intervals of estimates overlapping with 0  
564 (Figure S11). The low repeatability of latency measures between performance using different  
565 apparatuses may result from ceiling effects (e.g. individuals may solve an easy task with  
566 similar latencies but show greater variation when solving a more difficult problem) and floor  
567 effects (e.g. individuals may use the maximum time that is given in a trial to solve a more  
568 difficult problem but show variation for an easy task) [93,94]. Accordingly, the effects of  
569 internal or external variables on repeatability may be minimised by using binary measures  
570 such as success-or-failure (SUC). Our results indicate that certain types of measures (e.g.  
571 latency or the number of trials) used in some cognitive tasks are more sensitive to internal or  
572 external contextual variables than others and thus, provide less reliable measures of R.  
573 However, we suggest that moderator effects should be interpreted with caution, as constraints  
574 on our sample size prevented us from controlling for other fixed effects when revealing each  
575 moderator effect as well as potential interaction effects. Our approach of univariate testing

576 may therefore have been more liberal than a full model approach. While our results generally  
577 suggest that most moderators did not explain variation in the repeatability of inter-individual  
578 variation in cognitive performance across studies, these factors may still be important to  
579 consider when designing experiments for a particular species.

580

### 581 *General conclusion and future research*

582 To summarise, we report low to moderate estimates for the repeatability of cognitive  
583 performance, suggesting consistent individual differences over a range of cognitive tasks and  
584 taxa. Measurements of cognitive performance in a given task are therefore moderately  
585 consistent for individuals over time and can be studied much like other behavioral and  
586 morphological traits. Furthermore, different experimental paradigms that assess the same  
587 underlying cognitive capacity are reasonably concordant. This suggests that different  
588 approaches can be used to estimate the same underlying cognitive ability. Together, our  
589 results suggest that formally assessing individual variation in cognitive performance within  
590 populations could be a useful first step in research programs on the evolutionary biology of  
591 cognition.

592

593 While we attempt to understand the repeatability of cognitive performance, we acknowledge  
594 that this is an emerging and rapidly developing field. Accordingly, this study suffers some  
595 limitations, including a modest sample size (both for the number of studies included and for  
596 the number of subjects provided in each study) which reduces the robustness of the  
597 conclusions regarding the effect of potential moderators. Moreover, this study may also suffer  
598 some undetected bias in data collection, as the majority of data were obtained either from  
599 colleagues that presented at a workshop on the “Causes and consequences of individual  
600 variation in cognition” or researchers who work on individual differences known to the  
601 workshop participants. However, we argue that the inclusion of unpublished data is a useful  
602 approach to gaining a better representation of the true range of repeatabilities, given that we  
603 found published studies to provide higher R than unpublished studies. Future studies may  
604 therefore benefit from the growing body of literature on individual differences in cognition  
605 [81,82,95, Dougherty & Guillette in this issue]. Note that other studies collecting repeated  
606 measures from repetitions of a same test, or functionally-similar tests, could also offer  
607 valuable datasets, even when their aim is not the quantification of consistent individual  
608 differences. To facilitate future meta-analyses, we suggest that authors of such papers: (i)  
609 publish their datasets using the finest-grained information available (e.g. trial-by-trial instead

610 of aggregate values, such as proportion of correct choices or trials); (ii) include information  
611 on potential moderators (e.g. date of test, subject's origin) and other fixed effects (e.g. sex,  
612 age) that may need to be controlled for; and (iii) include and standardise the term 'cognitive  
613 repeatability' in their keywords.

614  
615 Future avenues for research may include: (1) studying the repeatability of reaction norms of  
616 cognitive performance (i.e. its plasticity [96,97] over gradients of interest, for example,  
617 deprivation level or housing conditions), so as to assess the generality of the individual  
618 differences that are captured by cognitive tasks across different environments and  
619 physiological states; and (2) partitioning the variance among and within individuals, by  
620 making use of multiple (>4) trials recorded for each individual [98]. By partitioning variance  
621 in cognitive performance at various hierarchical levels (within and between individuals) we  
622 may complement approaches that quantify variation at other levels (populations and species)  
623 and hence further our understanding of the evolution of cognition. This approach may  
624 provide a greater understanding of the factors that influence repeatability estimates, which are  
625 based on a ratio, and thus do not allow the separation of variance that is due to different  
626 phenotypes (among-individual) from those due to the plasticity in the response of each animal  
627 (within-individual). Separating these values could provide a way to focus on the portion of  
628 variance that is expected to be heritable, and to test hypotheses on the factors that affect  
629 variation within-individuals between repeated trials.

630 **References**

- 631 1. Shettleworth SJ. 2010 *Cognition, Evolution, and Behavior*. Oxford University Press.
- 632 2. van Horik J, Emery NJ. 2011 Evolution of cognition. *Wiley Interdiscip. Rev. Cogn. Sci.* **2**, 621–  
633 633.
- 634 3. Van Horik JO, Clayton NS, Emery NJ. 2012 *Convergent Evolution of Cognition in Corvids, Apes*  
635 *and Other Animals*.
- 636 4. MacLean EL *et al.* 2012 How does cognition evolve? Phylogenetic comparative psychology.  
637 *Anim. Cogn.* **15**, 223–238.
- 638 5. Thornton A, Isden J, Madden JR. 2014 Toward wild psychometrics: linking individual cognitive  
639 differences to fitness. *Behav. Ecol.* **25**, 1299–1301.
- 640 6. Cauchoix M, Chaine AS. 2016 How can we study the evolution of animal minds? *Front. Psychol.*  
641 **7**, 358.
- 642 7. Dukas R, Bernays EA. 2000 Learning improves growth rate in grasshoppers. *Proc. Natl. Acad.*  
643 *Sci. U. S. A.* **97**, 2637–2640.
- 644 8. Raine NE, Chittka L. 2008 The correlation of learning speed and natural foraging success in  
645 bumble-bees. *Proc R Soc Lond B Biol Sci.* **275**, 803–808.
- 646 9. Pasquier G, Grüter C. 2016 Individual learning performance and exploratory activity are linked to  
647 colony foraging success in a mass-recruiting ant. *Behav. Ecol.* **27**, 1702–1709.
- 648 10. Maille A, Schradin C. 2016 Survival is linked with reaction time and spatial memory in African  
649 striped mice. *Biol. Lett.* **12**. (doi:10.1098/rsbl.2016.0346)
- 650 11. Kotrschal A, Buechel SD, Zala SM, Corral-Lopez A, Penn DJ, Kolm N. 2015 Brain size affects  
651 female but not male survival under predation threat. *Ecol. Lett.* **18**, 646–652.
- 652 12. Keagy J, Savard J-F, Borgia G. 2009 Male satin bowerbird problem-solving ability predicts  
653 mating success. *Anim. Behav.* **78**, 809–817.
- 654 13. Cole EF, Morand-Ferron J, Hinks AE, Quinn JL. 2012 Cognitive ability influences reproductive  
655 life history variation in the wild. *Curr. Biol.* **22**, 1808–1812.
- 656 14. Cauchard L, Boogert NJ, Lefebvre L, Dubois F, Doligez B. 2013 Problem-solving performance is  
657 correlated with reproductive success in a wild bird population. *Anim. Behav.* **85**, 19–26.
- 658 15. Ashton BJ, Ridley AR, Edwards EK, Thornton A. 2018 Cognitive performance is linked to group  
659 size and affects fitness in Australian magpies. *Nature* **554**, 364–367.
- 660 16. Isden J, Panayi C, Dingle C, Madden J. 2013 Performance in cognitive and problem-solving tasks  
661 in male spotted bowerbirds does not correlate with mating success. *Anim. Behav.* **86**, 829–838.
- 662 17. Dunlap AS, Stephens DW. 2016 Reliability, uncertainty, and costs in the evolution of animal  
663 learning. *Curr. Opin. Behav. Sci* **12**, 73–79.
- 664 18. Mery F. 2013 Natural variation in learning and memory. *Curr. Opin. Neurobiol.* **23**, 52–56.
- 665 19. Kawecki TJ. 2009 Evolutionary ecology of learning: insights from fruit flies. *Popul. Ecol.* **52**,  
666 15–25.

- 667 20. Kotrschal A, Rogell B, Bundsen A, Svensson B, Zajitschek S, Brännström I, Immler S, Maklakov  
668 AA, Kolm N. 2013 Artificial selection on relative brain size in the guppy reveals costs and  
669 benefits of evolving a larger brain. *Curr. Biol.* **23**, 168–171.
- 670 21. Endler JA. 1986 *Natural Selection in the Wild*. Princeton University Press.
- 671 22. Dohm MR. 2002 Repeatability estimates do not always set an upper limit to heritability. *Funct.*  
672 *Ecol.* **16**, 273–280.
- 673 23. Edwards AWF, Falconer DS. 1982 Introduction to Quantitative Genetics. *Biometrics* **38**, 1128.
- 674 24. Wilson AJ. 2018 How should we interpret estimates of individual repeatability? *Evolution Letters*  
675 **2**, 4–8.
- 676 25. Dohm MR. 2002 Repeatability estimates do not always set an upper limit to heritability. *Funct.*  
677 *Ecol.* **16**, 273–280.
- 678 26. Griffin AS, Guillette LM, Healy SD. 2015 Cognition and personality: an analysis of an emerging  
679 field. *Trends Ecol. Evol.* **30**, 207–214.
- 680 27. Martin JGA, Réale D. 2008 Temperament, risk assessment and habituation to novelty in eastern  
681 chipmunks, *Tamias striatus*. *Anim. Behav.* **75**, 309–318.
- 682 28. Bell AM, Hankison SJ, Laskowski KL. 2009 The repeatability of behaviour: a meta-analysis.  
683 *Anim. Behav.* **77**, 771–783.
- 684 29. Dingemanse N, Réale D. 2005 Natural selection and animal personality. *Behaviour* **142**, 1159–  
685 1184.
- 686 30. Nicolaus M, Tinbergen JM, Bouwman KM, Michler SPM, Ubels R, Both C, Kempenaers B,  
687 Dingemanse NJ. 2012 Experimental evidence for adaptive personalities in a wild passerine bird.  
688 *Proc. R. Soc. B* **279**, 4885–4892.
- 689 31. Dingemanse NJ, Wolf M. 2010 Recent models for adaptive personality differences: a review.  
690 *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **365**, 3947–3958.
- 691 32. Dall SRX, Houston AI, McNamara JM. 2004 The behavioural ecology of personality: consistent  
692 individual differences from an adaptive perspective. *Ecol. Lett.* **7**, 734–739.
- 693 33. Hughes C, Adlam A, Happé F, Jackson J, Taylor A, Caspi A. 2000 Good test—retest reliability  
694 for standard and advanced false-belief tasks across a wide range of abilities. *J. Child Psychol.*  
695 *Psychiatry* **41**, 483–490.
- 696 34. Guenther A, Brust V. 2017 Individual consistency in multiple cognitive performance:  
697 behavioural versus cognitive syndromes. *Anim. Behav.* **130**, 119–131.
- 698 35. Brust V, Guenther A. 2017 Stability of the guinea pigs personality - cognition - linkage over  
699 time. *Behav. Processes* **134**, 4–11.
- 700 36. Gibelli J, Dubois F. 2016 Does personality affect the ability of individuals to track and respond to  
701 changing conditions? *Behav. Ecol.* **28**, 101–107.
- 702 37. Ashton BJ, Ridley AR, Edwards EK, Thornton A. 2018 Cognitive performance is linked to group  
703 size and affects fitness in Australian magpies. *Nature* **554**, 364–367.
- 704 38. Tello-Ramos MC, Branch CL, Pitera AM, Kozlovsky DY, Bridge ES, Pravosudov VV. 2018  
705 Memory in wild mountain chickadees from different elevations: comparing first-year birds with

- 706 older survivors. *Anim. Behav.* **137**, 149–160.
- 707 39. Chittka L, Dyer AG, Bock F, Dornhaus A. 2003 Psychophysics: bees trade off foraging speed for  
708 accuracy. *Nature* **424**, 388.
- 709 40. Dalesman S, Rendle A, Dall SRX. 2015 Habitat stability, predation risk and ‘memory  
710 syndromes’. *Sci. Rep.* **5**. (doi:10.1038/srep10538)
- 711 41. Morand-Ferron J, Cole EF, Quinn JL. 2016 Studying the evolutionary ecology of cognition in the  
712 wild: a review of practical and conceptual challenges. *Biol. Rev. Camb. Philos. Soc.* **91**, 367–389.
- 713 42. Rowe C, Healy SD. 2014 Measuring variation in cognition. *Behav. Ecol.* **25**, 1287–1292.
- 714 43. van Horik JO, Langley EJG, Whiteside MA, Laker PR, Beardsworth CE, Madden JR. 2018 Do  
715 detour tasks provide accurate assays of inhibitory control? *Proc. Biol. Sci.* **285**.  
716 (doi:10.1098/rspb.2018.0150)
- 717 44. Dingemanse NJ, Dochtermann NA. 2013 Quantifying individual variation in behaviour: mixed-  
718 effect modelling approaches. *J. Anim. Ecol.* **82**, 39–54.
- 719 45. Wilson AJ. 2018 How should we interpret estimates of individual repeatability? *Evolution Letters*  
720 **2**, 4–8.
- 721 46. Nakagawa S, Schielzeth H. 2010 Repeatability for Gaussian and non-Gaussian data: a practical  
722 guide for biologists. *Biol. Rev. Camb. Philos. Soc.* **85**, 935–956.
- 723 47. Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group. 2009 Preferred Reporting  
724 Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Med.* **6**,  
725 e1000097.
- 726 48. Griffin AS, Guillette LM, Healy SD. 2015 Cognition and personality: an analysis of an emerging  
727 field. *Trends Ecol. Evol.* **30**, 207–214.
- 728 49. Hughes C, Adlam A, Happé F, Jackson J, Taylor A, Caspi A. 2000 Good test—retest reliability  
729 for standard and advanced false-belief tasks across a wide range of abilities. *J. Child Psychol.*  
730 *Psychiatry* **41**, 483–490.
- 731 50. Duckworth AL, Kern ML. 2011 A meta-analysis of the convergent validity of self-control  
732 measures. *J. Res. Pers.* **45**, 259–268.
- 733 51. Rodríguez RL, Gloudeman MD. 2011 Estimating the repeatability of memories of captured prey  
734 formed by *Frontinella communis* spiders (Araneae: Linyphiidae). *Anim. Cogn.* **14**, 675–682.
- 735 52. Guenther A, Brust V. 2017 Individual consistency in multiple cognitive performance:  
736 behavioural versus cognitive syndromes. *Anim. Behav.* **130**, 119–131.
- 737 53. Schuster AC, Carl T, Foerster K. 2017 Repeatability and consistency of individual behaviour in  
738 juvenile and adult Eurasian harvest mice. *Naturwissenschaften* **104**, 10.
- 739 54. Schuster AC, Zimmermann U, Hauer C, Foerster K. 2017 A behavioural syndrome, but less  
740 evidence for a relationship with cognitive traits in a spatial orientation context. *Front. Zool.* **14**,  
741 19.
- 742 55. Shaw RC. 2017 Testing cognition in the wild: factors affecting performance and individual  
743 consistency in two measures of avian cognition. *Behav. Processes* **134**, 31–36.
- 744 56. Cole EF, Cram DL, Quinn JL. 2011 Individual variation in spontaneous problem-solving

- 745 performance among wild great tits. *Anim. Behav.* **81**, 491–498.
- 746 57. Koricheva J, Gurevitch J, Mengersen K. 2013 *Handbook of Meta-analysis in Ecology and*  
747 *Evolution*. Princeton University Press.
- 748 58. R Development Core Team. 2017 *R: A Language and Environment for Statistical Computing*.
- 749 59. Stoffel MA, Nakagawa S, Schielzeth H. 2017 rptR: repeatability estimation and variance  
750 decomposition by generalized linear mixed-effects models. *Methods Ecol. Evol.* **8**, 1639–1644.
- 751 60. Lessells CM, Boag PT. 1987 Unrepeatable Repeatabilities: A Common Mistake. *Auk* **104**, 116–  
752 121.
- 753 61. Holtmann B, Santos ESA, Lara CE, Nakagawa S. 2017 Personality-matching habitat choice,  
754 rather than behavioural plasticity, is a likely driver of a phenotype-environment covariance. *Proc*  
755 *R Soc Lond B Biol Sci.* **284**, 20170943.
- 756 62. Holtmann B, Lagisz M, Nakagawa S. 2016 Metabolic rates, and not hormone levels, are a likely  
757 mediator of between-individual differences in behaviour: a meta-analysis. *Funct. Ecol.* **31**, 685–  
758 696.
- 759 63. Wolak ME, Fairbairn DJ, Paulsen YR. 2011 Guidelines for estimating repeatability. *Methods*  
760 *Ecol. Evol.* **3**, 129–137.
- 761 64. Viechtbauer W. 2010 Conducting Meta-Analyses in R with the metafor Package. *J. Stat. Softw.*  
762 **36**. (doi:10.18637/jss.v036.i03)
- 763 65. Hinchliff CE *et al.* 2015 Synthesis of phylogeny and taxonomy into a comprehensive tree of life.  
764 *Proc. Natl. Acad. Sci. U. S. A.* **112**, 12764–12769.
- 765 66. Michonneau F, Brown JW, Winter D. 2016 rotl, an R package to interact with the Open Tree of  
766 Life data. (doi:10.7287/peerj.preprints.1471)
- 767 67. Nakagawa S, Schielzeth H. 2012 The mean strikes back: mean–variance relationships and  
768 heteroscedasticity. *Trends Ecol. Evol.* **27**, 474–475.
- 769 68. Benjamini Y, Hochberg Y. 1995 Controlling the false discovery rate: a practical and powerful  
770 approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **57**, 289–300.
- 771 69. Higgins JPT, Thompson SG. 2002 Quantifying heterogeneity in a meta-analysis. *Stat. Med.* **21**,  
772 1539–1558.
- 773 70. Nakagawa S, Santos ESA. 2012 Methodological issues and advances in biological meta-analysis.  
774 *Evol. Ecol.* **26**, 1253–1274.
- 775 71. Egger M, Davey Smith G, Schneider M, Minder C. 1997 Bias in meta-analysis detected by a  
776 simple, graphical test. *BMJ* **315**, 629–634.
- 777 72. Nakagawa S, Santos ESA. 2012 Methodological issues and advances in biological meta-analysis.  
778 *Evol. Ecol.* **26**, 1253–1274.
- 779 73. Moreno SG, Sutton AJ, Ades AE, Stanley TD, Abrams KR, Peters JL, Cooper NJ. 2009  
780 Assessment of regression-based methods to adjust for publication bias through a comprehensive  
781 simulation study. *BMC Med. Res. Methodol.* **9**, 2.
- 782 74. Nakagawa S, Noble DWA, Senior AM, Lagisz M. 2017 Meta-evaluation of meta-analysis: ten  
783 appraisal questions for biologists. *BMC Biol.* **15**, 18.



- 784 75. Bell AM, Hankison SJ, Laskowski KL. 2009 The repeatability of behaviour: a meta-analysis.  
785 *Anim. Behav.* **77**, 771–783.
- 786 76. Croston R, Branch CL, Kozlovsky DY, Dukas R, Pravosudov VV. 2015 Heritability and the  
787 evolution of cognitive traits. *Behav. Ecol.* **26**, 1447–1459.
- 788 77. Kotrschal A, Rogell B, Bundsen A, Svensson B, Zajitschek S, Brännström I, Immler S, Maklakov  
789 AA, Kolm N. 2013 Artificial selection on relative brain size in the guppy reveals costs and  
790 benefits of evolving a larger brain. *Curr. Biol.* **23**, 168–171.
- 791 78. Burger JMS, Kolss M, Pont J, Kawecki TJ. 2008 Learning ability and longevity: a symmetrical  
792 evolutionary trade-off in *Drosophila*. *Evolution* **62**, 1294–1304.
- 793 79. Snell-Rood EC, Davidowitz G, Papaj DR. 2011 Reproductive tradeoffs of learning in a butterfly.  
794 *Behav. Ecol.* **22**, 291–302.
- 795 80. Tryon RC. 1940 Studies in individual differences in maze ability. VII. The specific components  
796 of maze ability, and a general theory of psychological components. *J. Comp. Psychol.* **30**, 283–  
797 335.
- 798 81. Thornton A, Lukas D. 2012 Individual variation in cognitive performance: developmental and  
799 evolutionary perspectives. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **367**, 2773–2783.
- 800 82. Rowe C, Healy SD. 2014 Measuring variation in cognition. *Behav. Ecol.* **25**, 1287–1292.
- 801 83. Wäckers FL, Lewis WJ. 1999 A comparison of color-, shape- and pattern-learning by the  
802 hymenopteran parasitoid *Microplitis croceipes*. *J Comp. Physiol. A* **184**, 387–393.
- 803 84. Aronsson M, Gamberale-Stille G. 2008 Domestic chicks primarily attend to colour, not pattern,  
804 when learning an aposematic coloration. *Anim. Behav.* **75**, 417–423.
- 805 85. O’Hara M, Huber L, Gajdon GK. 2015 The advantage of objects over images in discrimination  
806 and reversal learning by kea, *Nestor notabilis*. *Anim. Behav.* **101**, 51–60.
- 807 86. Chow PKY, Leaver LA, Wang M, Lea SEG. 2017 Touch screen assays of behavioural flexibility  
808 and error characteristics in Eastern grey squirrels (*Sciurus carolinensis*). *Anim. Cogn.* **20**, 459–  
809 471.
- 810 87. Biro PA, Stamps JA. 2015 Using repeatability to study physiological and behavioural traits:  
811 ignore time-related change at your peril. *Anim. Behav.* **105**, 223–230.
- 812 88. Ono M, Kawai R, Horikoshi T, Yasuoka T, Sakakibara M. 2002 Associative learning acquisition  
813 and retention depends on developmental stage in *Lymnaea stagnalis*. *Neurobiol. Learn. Mem.* **78**,  
814 53–64.
- 815 89. Ushitani T, Perry CJ, Cheng K, Barron AB. 2016 Accelerated behavioural development changes  
816 fine-scale search behaviour and spatial memory in honey bees (*Apis mellifera* L.). *J. Exp. Biol.*  
817 **219**, 412–418.
- 818 90. Jonasson Z. 2005 Meta-analysis of sex differences in rodent models of learning and memory: a  
819 review of behavioral and biological data. *Neurosci. Biobehav. Rev.* **28**, 811–825.
- 820 91. Vallortigara G. 1996 Learning of colour and position cues in domestic chicks: Males are better at  
821 position, females at colour. *Behav. Processes* **36**, 289–296.
- 822 92. Laland KN, Reader SM. 1999 Foraging innovation in the guppy. *Anim. Behav.* **57**, 331–340.

- 823 93. Griffin AS, Guez D. 2014 Innovation and problem solving: a review of common mechanisms.  
824 *Behav. Processes* **109 Pt B**, 121–134.
- 825 94. van Horik JO, Madden JR. 2016 A problem with problem solving: motivational traits, but not  
826 cognition, predict success on novel operant foraging tasks. *Anim. Behav.* **114**, 189–198.
- 827 95. Morand-Ferron J, Quinn JL. 2015 The evolution of cognition in natural populations. *Trends*  
828 *Cogn. Sci.* **19**, 235–237.
- 829 96. Dingemanse NJ, Dochtermann NA. 2013 Quantifying individual variation in behaviour: mixed-  
830 effect modelling approaches. *J. Anim. Ecol.* **82**, 39–54.
- 831 97. Martin JGA, Nussey DH, Wilson AJ, Réale D. 2011 Measuring individual differences in reaction  
832 norms in field and experimental studies: a power analysis of random regression models. *Methods*  
833 *Ecol. Evol.* **2**, 362–374.
- 834 98. van de Pol M, Wright J. 2009 A simple method for distinguishing within- versus between-subject  
835 effects using mixed models. *Anim. Behav.* **77**, 753–758.
- 836 88. Biro PA, Stamps JA. 2015 Using repeatability to study physiological and behavioural traits:  
837 ignore time-related change at your peril. *Anim. Behav.* **105**, 223–230.

838

839 **Figure and table captions**

840 Figure 1: Temporal repeatability R (unadjusted) and 95% bootstrapped confidence intervals  
841 for each dataset. Y-axis provides information about first author, species name, the type of  
842 cognitive task and the type of cognitive performance measurement. Cognitive performance  
843 measurement was the quantification of a cognitive process using: accuracy such as proportion  
844 correct (ACC); the number of trials to reach a learning criterion (TTC); success-or-failure  
845 binary outcome (SUC); latency (LAT); normalised performance scores (NOR); the number of  
846 correct trials or errors over a fixed number of trials (NBT). The type of cognitive task include  
847 mechanical problem solving (PS); discriminative learning (DL); reversal learning (RL);  
848 inhibition (IN); memory (ME); use of human cue (HC); external attention (EA); internal  
849 attention (IA); learning (LE); physical cognition (PC) that includes visual exclusion  
850 performance; auditory exclusion performance and object permanence; social learning (SL),  
851 spatial orientation learning (SOL), spatial recognition (SR) and lexical fluency (LF).

852

853 Figure 2: Contextual repeatability R (unadjusted) and 95% bootstrapped confidence intervals  
854 for each dataset. Y-axis presents first author, species name, the type of cognitive task and the  
855 type of cognitive performance measurement. Cognitive measurement is used to quantify a  
856 cognitive process using: accuracy such as proportion correct (ACC); the number of trials to  
857 reach a learning criterion (TTC); success-or-failure binary outcome (SUC); latency (LAT);  
858 normalised performance scores (NOR); the number of correct trials or errors over a fixed

859 number of trials (NBT). The types of cognitive task include mechanical problem solving (PS);  
860 discriminative learning (DL); reversal learning (RL); inhibition (IN); memory (ME); use of  
861 human cue (HC); external attention (EA); internal attention (IA); learning (LE); physical  
862 cognition (PC) that includes visual exclusion performance; auditory exclusion performance  
863 and object permanence; social learning (SL), spatial orientation learning (SOL), spatial  
864 recognition (SR) and lexical fluency (LF).

865

866 Figure 3: Meta-analytic mean estimates of repeatability (R) for temporal and contextual  
867 repeatability including unadjusted, adjusted for test order and adjusted for test order plus  
868 individual determinants (sex and/or age). We present posterior means and 95% confidence  
869 intervals (CIs) of meta-analyses obtained from linear mixed-effects models (LMMs). All  
870 estimates are back-transformed into repeatability (R).

871

872 Table 1: Summary results from meta-analytic model: mean estimates, upper and lower  
873 confidence interval, sample size (total number of R value considered in the analysis), Egger's  
874 regression significance (P-value), total heterogeneity, partial heterogeneity due to the  
875 laboratory, species and experiment.

876 Table 2: Summary of meta-regression models. Conditional  $R^2$  and significance (P-values from  
877 omnibus test) of each moderator from the 7 univariate meta regressions are presented.

878 *Data accessibility.* We provide access to the information of general methods (ESM) and  
879 primary data (<https://doi.org/10.6084/m9.figshare.6431549.v1>).

880 *Ethics.* All studies complied with local ethics regulations as listed in the associated  
881 publication. Completely unpublished data provide this information in the online methods.

882 *Competing interests.* All authors declare there is no competing interests.

883 *Funding.* PKYC is supported by Japan Society for the Promotion of Science (PE1801); JOvH  
884 was funded by an ERC consolidator grant (616474). MC and this research was supported by a  
885 grant from the Human Frontier Science Program to ASC and JM-F (RGP0006/2015).

886

887

888