

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

2

3 Cauchoix M^{1,2*}, Chow PKY^{3*}, van Horik JO^{3*}, Atance CM⁴, Barbeau EJ⁵, Barragan-Jason G²,
4 Bize P⁶, Boussard A⁷, Buechel SD⁷, Cabirol A⁸, Cauchard L⁹, Claidière N¹⁰, Dalesman S¹¹,
5 Devaud JM⁸, Didic M¹², Doligez B¹³, Fagot J¹⁰, Fichtel C¹⁴, Henke-von der Malsburg J¹⁴,
6 Hermer E¹⁵, Huber L¹⁶, Huebner F¹⁴, Kappeler PM^{14,17}, Klein S⁸, Langbein J¹⁸, Langley EJG³,
7 Lea SEG³, Lihoreau M⁸, Lovlie H¹⁹, Matzel LD²⁰, Nakagawa S²¹, Nawroth C¹⁸, Oesterwind
8 S²², Sauce B²⁰, Smith E²³, Sorato E¹⁹, Tebbich S²⁴, Wallis LJ^{16,25}, Whiteside MA³, Wilkinson
9 A²³, Chaine AS^{1,2§}, Morand-Ferron J^{15§}.

10

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

13 ²Institute for Advanced Studies in Toulouse, 21 allée de Brienne, 31015, Toulouse, France

14 ³Centre for Research in Animal Behaviour, Psychology, University of Exeter, UK.

15 ⁴School of Psychology, University of Ottawa, Ottawa, Canada

16 ⁵Centre de recherche Cerveau et Cognition, UPS-UMR5549, Toulouse, France

17 ⁶Institute of Biological and Environmental Sciences, University of Aberdeen, Aberdeen, U.K.

18 ⁷Department of Zoology/Ethology, Stockholm University, Svante Arrheniusväg 18B, 10691 Stockholm, Sweden

19

20 ⁸Research Center on Animal Cognition (CRCA), Center for Integrative Biology (CBI); CNRS, University Paul
21 Sabatier, Toulouse

22 ⁹Département de Sciences Biologiques, Université de Montréal, Montreal, QC, Canada

23 ¹⁰Aix Marseille University, CNRS, LPC, Marseille, France

24 ¹¹Institute of Biological, Environmental and Rural Sciences, Aberystwyth University, Aberystwyth, U.K.

25 ¹²AP-HM Timone & Institut de Neurosciences des Systèmes, Marseille, France

26 ¹³CNRS UMR 5558, Université Lyon 1, Department of Biometry and Evolutionary Biology, France

27 ¹⁴Behavioral Ecology & Sociobiology Unit, German Primate Center, Göttingen, Germany

28 ¹⁵Department of Biology, University of Ottawa, Ottawa, Canada.

29 ¹⁶Comparative Cognition, Messerli Research Institute, University of Veterinary Medicine Vienna, Medical
30 University of Vienna, University of Vienna, Vienna, Austria

31 ¹⁷Department of Sociobiology/ Anthropology, University of Göttingen, Göttingen, Germany

32 ¹⁸Institute of Behavioural Physiology, Leibniz Institute for Farm Animal Biology, Dummerstorf, Germany

33 ¹⁹IFM Biology, Linköping University, 58183 Linköping, Sweden

34 ²⁰Department of Psychology, Rutgers University, Piscataway, USA

35 ²¹Evolution & Ecology Research Centre, School of Biological, Earth & Environmental Sciences, University of
36 New South Wales, Sydney, NSW 2052, Australia

37 ²²Faculty of Agricultural and Environmental Sciences, University of Rostock, Rostock, Germany

38 ²³School of Life Sciences, University of Lincoln, Lincoln, UK.

39 ²⁴Department of Behavioural Biology, University of Vienna, Austria

40 ²⁵Department of Ethology, Eötvös Loránd University, Budapest, Hungary

41

42 *Shared first authorship listed alphabetically

43 §Shared senior authorship listed alphabetically

44

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

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

52

53

54 **ABSTRACT**

55 Selection acts on heritable individual variation in behaviours. Both behavioural and cognitive
56 processes play important roles in mediating an individual's interactions with their environment.
57 Yet, while there is a vast literature on repeatable individual differences in behaviour, relatively
58 little is known about the repeatability of cognitive performance. To further our understanding
59 of the evolution of cognition we gathered 44 datasets on individual performances of 25 species
60 and used meta-analysis to evaluate whether cognitive performance is repeatable across six
61 animal classes. We assessed repeatability (R) in performance (1) on the same task presented at
62 different time intervals (temporal repeatability), and (2) on different tasks that measure the same
63 putative cognitive ability (contextual repeatability). We also addressed whether R estimates are
64 influenced by seven extrinsic factors (moderators): type of cognitive task, type of measurement,
65 delay between tasks, origin of the subjects, experimental context, taxonomic class and if the R
66 value was published or unpublished. We found support for both temporal and contextual
67 repeatability of individual variation in cognitive performance, with significant mean R
68 estimates ranging between 0.15 and 0.28. R estimates were mostly influenced by the type of
69 cognitive performance measures and the fact that R values was published, none of the other
70 moderators showed consistent and significant impacts on repeatability estimates. Our overall
71 findings highlight the widespread occurrence of consistent inter-individual variation in
72 cognition which, like behaviour, may have fitness implications.

73

74 *Keywords:* cognitive repeatability; consistency; evolutionary biology of cognition; individual
75 differences; learning; memory; attention.

76

77

78 INTRODUCTION

79

80 Cognition has been broadly defined as the acquisition, processing, storage and use of
81 information [1], and hence plays an important role in mediating how animals behave and
82 interact with their environment. While comparative studies have broadened our understanding
83 of how socio-ecological selection pressures shape cognitive evolution [2–4], relatively little is
84 known about the adaptive significance of inter-individual variation of cognitive abilities [5,6].
85 There is however some evidence that learning may be under selection if it influences fitness [6-
86 19]. Opportunities to learn have been linked to increased growth rate [7], and individual
87 learning speed can correlate with foraging success [8,9]. Greater cognitive capacities may allow
88 individuals to better detect and evade predators [10,11] and may also influence their
89 reproductive success [12–15]; but see [16]. Finally, rapid evolutionary change in learning
90 abilities have also been shown by experimentally manipulating environmental conditions,
91 revealing trade-offs between fitness benefits and costs to learning [17–20]. Accordingly, we
92 might expect selection to act on individual differences in cognitive ability in other species and
93 contexts.

94

95 As selection acts on variation, a fundamental prerequisite to understanding the evolution of
96 cognition in extant populations requires an assessment of individual variation in cognitive traits
97 [21]. The approach most commonly used in evolutionary and ecological studies to estimate
98 consistent among-individual variation has its origin in quantitative genetics [22,23]. This
99 approach compares the variation in two or more measures of the same individual, with variation
100 in the same trait across all individuals to distinguish between variation due to “noise” and
101 variation among individuals. The amount of variation explained by inter-individual variation
102 relative to intra-individual variation is termed the “intraclass correlation coefficient” or
103 “repeatability” (R). Repeatability coefficients are often used to estimate the upper limit of
104 heritability [23] but see [22], and thus quantifying repeatability is a useful first step in
105 evolutionary studies of traits [24].

106

107

108 Assessing the repeatability of behavioural or cognitive traits is, however, challenging, because
109 the context of measurement can influence the behaviour of animals, and thus, the value
110 recorded. Contextual variation can come from the internal state of the organism (e.g. hunger,
111 circadian cycle, recent interactions, stress) or the external environment, which may differ

112 between trials [25]. Moreover, behavioural and cognitive measures may suffer further variation
113 between measures as experience with one type of measure or test can influence subsequent
114 measures via processes such as learning and memory [26]. While this issue has been recognised
115 and discussed in recent research on animal personality [27], it may be particularly relevant
116 when assaying the repeatability of cognitive traits. Consequently, we might therefore expect
117 higher within-individual variation in behavioural or cognitive measures compared with
118 morphological or physiological measures, due to greater differences in the context (internal or
119 external) of repeated sampling.

120
121 Research on animal personality has provided a broad understanding that individual differences
122 in behaviour are repeatable (average $R = 0.37$) across time and contexts [28], hence revealing
123 an important platform for selection to act on [29–32]. Yet, relatively little is known about the
124 stability of inter-individual variation in cognitive traits, such as those associated with learning
125 and memory [26]. Some examples of repeatability estimates suggest that children show good
126 test–retest reliability on false-belief tasks used to assess theory-of-mind [26,33]. Consistent
127 individual differences in performance on cognitive tasks have also been documented in a few
128 non-human animals, such as guinea pigs, *Cavia aperea f. porcellus* [34,35], zebra finch,
129 *Taenopigya guttata* [36], Australian magpies, *Gymnorhina tibicen* [37], mountain chickadees,
130 *Poecile gambeli* [38], bumblebees, *Bombus terrestris* [39] and snails, *Lymnaea stagnalis* [40].
131 While the paucity of repeatability measures of cognitive performance may stem from the
132 recency of interest in the evolutionary ecology of cognitive traits [41,42], it may also suggest
133 that it is difficult to accurately capture repeatable measures of cognitive ability [43].

134
135 Recent advances in analytical techniques, such as the use of mixed-effect models, have
136 facilitated the assessment of repeatability of behavioural traits, by accounting for the potential
137 confounding effects of both internal and external contextual variations [44,45]. Such
138 approaches can help provide more accurate estimates of repeatability of cognitive traits and
139 could provide new insights to the influence of internal and external factors on cognitive
140 performance. For example, we can now explicitly address the effect of time, or an individual's
141 condition, on the repeatability of traits of interest such as learning performance. Likewise, we
142 can examine the effect of external factors, for example by modeling the environment (e.g. group
143 size at testing) or the type of test employed (e.g. spatial vs. colour cues in associative learning).
144 Adopting these methods (i.e. adjusted repeatability [46]) could therefore facilitate studies that

145 generate repeatability estimates of cognitive performance and provide greater clarity into the
146 sources of variation in measures of cognition in this rapidly expanding field.

147
148 In this study, we use meta-analysis to (1) estimate average repeatability of cognitive
149 performance across different taxa, and (2) discuss the implications of these results for how we
150 measure cognition and the importance of internal and external factors on the repeatability of
151 cognition. To do this we assessed individual performances from 14 different cognitive tasks
152 from 25 species of six animal classes. For each of the 14 tasks, we assessed multiple
153 performance measures, such as trials to criterion or success-or-failure for the same task. We
154 then assessed *temporal repeatability* by comparing individual performances on multiple
155 exposures of the same task, and *contextual repeatability* by comparing individual performances
156 on different tasks that measure the same putative cognitive ability. We then used meta-analysis
157 to investigate whether there are general across-taxa patterns of repeatability for different tasks
158 and which factors (type of cognitive performance measurement, type of cognitive task, delay
159 between tasks, origin of the subjects, experimental context, taxonomic class, and whether the
160 R value was published or unpublished) might influence the repeatability of cognitive
161 performance.

162

163

164

165

166 **METHODS**

167

168 **Data collection**

169 We followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses
170 (PRISMA) approach for the collation of the datasets used in the current study [47]. We first
171 collected published repeatability estimates of cognitive performance (Figure S1). We did not
172 include studies reporting inter-class correlations (Pearson or Spearman) between cognitive
173 performances on tasks measuring different cognitive abilities (i.e., general intelligence) as we
174 considered these outside the scope of this meta-analysis. Although we acknowledge that results
175 from the literature on test-retest [48,49] or convergent validity [50] in psychology would be
176 relevant to compare with the present study, we also considered them beyond the scope of this
177 paper as their inclusion would have led to a heavy bias towards studies on humans. We only
178 found 6 publications reporting repeatability of cognitive performance (R) in 6 different species:

179 1 arachnid [51], 2 mammals [52–54] and 3 birds [15,55,56], with a sample size ranging from
180 15 to 347 (mean: 54.66, median: 33) and number of repeated tests varying from 2 to 4 (mean:
181 2.5, median: 2).

182
183 To complement our data set from published studies, we used an ‘individual-patient-data’ meta-
184 analysis approach commonly used in medical research [57] in which effect sizes are extracted
185 using the same analysis on primary data [57]. We invited participants from a workshop on the
186 ‘Causes and consequences of individual variation in cognitive ability’ (36 people), as well as
187 25 colleagues working on individual differences in cognition, to contribute primary datasets of
188 repeated measurements of cognitive performance. From this approach, we assembled 38
189 primary datasets from unpublished (9 datasets: 6 studies were fully unpublished while 3 had
190 similar methods published from the same laboratory group) or published sources (29 datasets
191 but the data needed to calculate repeatability were not provided in the publications), from which
192 we could compute repeatability using the same analytical methods (Figure S1, see shared
193 repository link). These datasets comprised 20 different species of mammals (humans included),
194 insects, molluscs, reptiles and birds (Table S1 and Table S2). Details about subjects,
195 experimental context and cognitive tasks for each dataset can be found in electronic
196 supplementary material (ESM methods).

197
198 Each dataset included 4 – 375 individuals (mean: 46.6, median: 29), that performed 2 – 80
199 (mean: 7.9, median: 2) repetitions of tests targeting the same cognitive process, either by
200 conducting the same task presented at different points in time (*temporal repeatability*, see Table
201 S1), or different tasks aimed at testing the same underlying cognitive process but using a
202 different protocol (*contextual repeatability*, see Table S2). Tasks considered to assess
203 contextual repeatability differed by stimulus dimension (e.g. spatial vs. colour reversal learning
204 in Cauchoix- great tit dataset), sensory modality (e.g. visual vs. olfactory discrimination in
205 Henke- v.d. Malsburg -microcebus dataset), change in experimental apparatus (e.g. colour
206 discrimination on touch screen and on solid objects in Chow-squirrel lab dataset) or could be a
207 different task designed to measure the same cognitive process (i.e. Mouse Stroop Test and the
208 Dual Radial Arm Maze to measure external attention in Matzel-attention mice dataset).

209 210 **Repeatability analysis for primary data**

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

- 213 (1) We first transformed cognitive variables if necessary to meet assumptions of normality.
- 214 (2) To understand if taking into account the number of repetitions, test order, and/or an
215 individual's sex and age (hereafter, individual determinants) played a role in repeatability of
216 cognitive performances, we then computed 3 types of repeatability values with a mixed-effects
217 model approach using the appropriate link function in the 'rptR' package [59]. Specifically, we
218 calculated unadjusted repeatability (R), repeatability adjusted for test order (R_n), and
219 repeatability adjusted for test order and individual determinants (R_{ni}) and we calculated each
220 of these metrics for *temporal* and *contextual* repeatability separately.
- 221 (3) For cases with unadjusted R close to 0 (< 0.005), we computed the R estimate using a least
222 squares ANOVA approach as advised in [60–62] using the 'ICC' package [63].
- 223 (4) Finally, we removed R estimates from further analyses when residuals were not normal or
224 overdispersed (Poisson distribution) and data could not be transformed to achieve normality.
225 See ESM general methods for more details.

226

227 **Meta-analysis and meta-regression**

228 We collated the 178 R values computed from primary data with the 35 from published R values,
229 to obtain a total of 213 estimates of cognitive repeatability. We didn't compute repeatability de
230 novo for published study as the statistics used in these papers are the same or similar to the one
231 we used here for primary data (e.g. mixed-model approach with or without 'rptR' package). We
232 then used a meta-analytic approach to examine average across species repeatability of cognitive
233 performance. This approach allowed us to: (1) take into account sample size and number of
234 repeated measure associated with each R value in the estimation of average cognitive
235 repeatability, (2) control for repeated samples (i.e., avoid pseudoreplication) of the same species
236 (taxonomic bias), the same laboratory group (i.e., same senior author; observer bias) or the
237 same experiment (measurement bias) by including these factors as random effects, and (3) ask
238 whether other specific factors (fixed effects called "moderators" in meta-analysis, see below)
239 could explain the variation in repeatability of cognitive tests.

240

241 For each of the 6 type of R analysis (unadjusted temporal R, adjusted temporal R for test order,
242 adjusted temporal R for test order and individual determinants, unadjusted contextual R,
243 adjusted contextual R for test order, adjusted contextual R for test order and individual
244 determinants), we performed 3 different multilevel meta-analyses, by fitting Linear Mixed
245 Models (LMMs) using the 'metafor' package [64]: (1) a standard meta-analytic model
246 (intercept-only model) to estimate the overall mean effect size, (2) 7 univariate (multilevel)

247 meta-regression models to independently test the significance of each moderator. For each
248 model, we used standardized (Fisher's Z transformed) R values as the response variable.
249 Finally, we conducted (3) a type of Egger's regression to test for selection bias.

250

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

256

257 In meta-regression models, we accounted for variance in repeatability of cognitive traits by
258 adding both fixed and random effects. We accounted for variation in repeatability related to
259 fixed effects by including moderators. We considered 7 moderators (detailed in ESM general
260 methods and Figure 1 and 2 captions): type of cognitive performance measurement (e.g. success
261 or failure, latency, the number of trials before reaching a learning criterion); type of cognitive
262 task (e.g. reversal learning, discrimination learning); median delay between tests; experimental
263 context (conducted in the wild or in captivity); the origin of subjects (wild or hand raised),
264 taxonomic class, and if the R value was published or unpublished. In addition to fixed effect
265 moderators, we also took into account non-independence of data by including a series of
266 random effects. We included random effects for species (multiple datasets from the same
267 species), laboratory group (experiments conducted by the same PI), and experiment
268 (experiments on the same subjects; see ESM general methods for more details).

269

270 We controlled for the possibility that phylogenetic history influences the repeatability of
271 cognitive abilities (i.e. similar species have more similar repeatability of cognitive abilities) by
272 using a covariance matrix based on an order-level phylogenetic tree (using Open Tree of Life
273 [65] and "rotl" R package [66]) but only in the intercept only model as meta-regression models
274 failed to converge with this additional information. We ran the intercept only meta-analysis
275 with and without controlling for the effect of phylogeny and found that phylogenetic
276 relationships had negligible effects on average repeatability of cognitive abilities (Table S5),
277 justifying its exclusion in subsequent meta-regression models.

278

279 For meta-regressions, we report conditional R² (sensu [67]) which quantifies the proportion of
280 variance explained by fixed (moderators) and random effects along with p-values from omnibus

281 tests [64] which test the significance of multiple moderator effects. When omnibus tests were
282 significant ($p < 0.05$) we ran the same meta-regression model but without the intercept to
283 compute and plot beta coefficients associated with each level of the moderator (Figure S10 and
284 S11), and performed multiple pairwise comparisons to estimate statistical differences between
285 all combinations of moderator levels. We corrected for multiple comparisons using a false
286 discovery rate adjustment of p-values [68].

287
288 We assessed the extent of variation among effect sizes in each meta-analytic model (intercept
289 only) by calculating heterogeneities (I^2). Along with the overall heterogeneity (I^2_{total}), which
290 represents between-study variance divided by the total variance [69], we also provide estimates
291 of heterogeneity for each random factor (species, laboratory and experiment) following [70]. I^2
292 values of 25%, 50% and 75% are generally considered to be low, moderate and high levels of
293 heterogeneity, respectively [69].

294
295 Finally, we statistically tested for selection bias in the dataset by conducting a type of Egger's
296 regression [71]. Given that our effect sizes were not independent from each other, we employed
297 a mixed-model version of Egger's regression using the full models (7 moderators as fixed
298 effects) with the sampling standard errors (SE) of each effect size as a moderator [72,73]; a
299 regression slope of the SE significantly different from zero indicates selection bias [71]. Such
300 a significant effect usually means that large effect sizes with large sampling variance (small
301 sample size) are more prevalent than expected, potentially overestimating the overall effect size
302 (i.e., R).

303

304 **RESULTS**

305 *Dataset summary*

306 Repeatability estimates computed from primary data are presented together with published R
307 values in Table S1 for temporal repeatability and Table S2 for contextual repeatability. For
308 temporal repeatability, we used 22 studies on 15 species in which 4 to 375 (mean: 56.31,
309 median: 40) individuals performed a median of 2, 95%CI [1.91, 2.11] repeated tests, leading to
310 a total of 106 repeatability analyses (40 R; 40 Rn; and 26 Rni). For contextual repeatability, we
311 used 27 studies on 20 species in which 4 to 297 (mean: 41, median: 24) individuals performed
312 a median of 2, 95%CI [1.80, 2.15] repeated tests, leading to a total of 107 repeatability analysis
313 (38 R; 32 Rn; and 37 Rni).

314

315 ***Repeatabilities for individual studies***

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

324

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

326 We first used meta-analysis (intercept-only) models to compute mean estimates of cognitive
327 repeatability while taking into account variation in sample size and repetition number between
328 studies. Intercept-only models reveal significant low to moderate [0.15 - 0.28] mean estimates
329 of cognitive repeatability across analyses (Table 1, Figure 3). Performing the same analysis
330 with or without controlling for phylogenetic history suggests that class-level phylogenetic
331 relationships had little influence on mean cognitive repeatability estimates (Table S4).

332

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

342

343 We investigated whether our meta-analysis model showed any bias in data publication or
344 selection using a type of Egger's regression. Egger's regressions suggest significant bias for
345 unadjusted temporal R. Such bias is probably related to the high number of low sample size
346 studies. To further evaluate the robustness of our mean estimates, we ran a sensitivity analysis
347 using a "leave one out procedure" (ESM general methods) in which we computed mean
348 estimates by removing a single R value for each R value in the dataset and generating a

349 distribution of mean estimates. The distribution of “leave one out” mean estimates were
350 concentrated around the original mean estimate, which suggests that meta-analytic results are
351 not driven by one particular R value (Figure S10). Finally, we assessed whether mean estimates
352 obtained for each type of R analysis was significantly different from each other using multiple
353 t-test comparisons. We found that adjusted temporal R for test order was significantly lower
354 than other types of R analyses before correcting for multiple comparisons (Table S5). However,
355 we found no significant differences after correcting for multiple comparisons for all
356 combinations of R analyses.

357

358 *Meta-regression: effects of moderators*

359 To better understand the factors that influence heterogeneity of repeatability, we included the
360 type of cognitive performance measurement, the type of cognitive task, median delay between
361 repetitions, experimental context, origin of the subjects, taxonomic class, and publication status
362 as moderators in our models of repeatability. Effects of those factors on raw R values can be
363 inspected visually in Figures S3-9. However, to assess the effects of these factors while
364 accounting for variation in sample size and repetition number between studies, meta-analytical
365 tools are necessary. The total number of repeatability values compiled for each type of R
366 analysis (Table 1) was not sufficient to run a full model to assess the effects of all 7 moderators
367 together. We therefore ran 7 independent univariate (multilevel) meta-regression models, which
368 revealed that measures of cognitive performance significantly influenced all types of R
369 analyses, except for temporal unadjusted values (Table 2), and accounted for 14 to 100% of the
370 variance (R^2_c). The investigation of beta coefficients associated with each type of cognitive
371 measurement (Figure S11) suggests that normalized index (score computed specifically for the
372 study e.g. Matzel et al. dataset) and success measures are significantly more repeatable for
373 contextual R_{ni} estimates than other types of R analyses. However, as this pattern is not observed
374 for other types of R analyses, results should be interpreted with caution. The publication of R
375 values also significantly influenced contextual repeatability and accounted for 24 to 70% of the
376 variance (Table 2), with published R values being significantly higher than R computed from
377 primary data (Figure S12).

378

379 We found that the type of cognitive task, median delay between tasks, experimental context,
380 the origin of the subjects or taxonomic class did not show consistently significant effects across
381 different types of R analyses. The significant effect of cognitive task type on unadjusted
382 contextual R should be interpreted cautiously as it is present only for one type of R analysis and

383 is thus probably not robust (Table 1 and Figure 1). The same is also true for the marginally
384 significant effect of median delay between tasks; its positive beta coefficient (0.06, see also
385 Figure S3) suggests that repeatability increased with the delay between tests. This finding could
386 be driven by high R values from the study by Barbeau et al. in humans (Table S1) despite a
387 very long median delay between trials (540 days). Indeed, the p-value associated to median
388 delay became non-significant when running the same meta-regression without those data.

389

390

391 **DISCUSSION**

392 We aimed to explore the repeatability of cognitive performance across six animal classes. We
393 examined repeatability by assessing whether inter-individual variation in cognitive
394 performance was consistent on the same task across two or more points in time (i.e. temporal
395 repeatability) or whether performances were consistent across different tasks that are designed
396 to capture the same cognitive process (i.e. contextual repeatability). Overall, our meta-analysis
397 revealed robust and significant low to moderate repeatability of cognitive performance ($R =$
398 $[0.15-0.28]$). We found that the type of cognitive performance measurement (e.g. the number
399 of trials to reach a criterion, latency) affected most estimates of repeatabilities while the type of
400 cognitive task (e.g. reversal learning, discrimination learning, mechanical problem solving),
401 delay between task repetitions, the origin of animals (wild/wild-caught or laboratory-
402 raised/hand-raised), experimental context (in the wild or laboratory), taxonomic class, and
403 origin of R values (published vs. primary data) did not consistently show significant effects on
404 R estimates.

405

406 *Are measures of cognition repeatable?*

407

408 High plasticity of cognitive processes could have been expected to result in very low or null
409 estimates of repeatability. Yet, we found a significant, but low average R estimate for
410 unadjusted temporal repeatability of cognitive performance ($R = 0.15$). Our highest temporal
411 repeatability estimate adjusted for test order and individual determinants attained $R = 0.28$.
412 Although this estimate remains lower than that observed for animal personality ($R = 0.37$) [75],
413 our findings suggest that inter-individual variation in performance on the same cognitive task
414 is moderately consistent across time in a wide range of taxa. This result is particularly striking
415 because internal and external influences on task performance are unlikely to be identical
416 between trials; such influences should inflate intra-individual variation between trials, and

417 therefore reduce R. The results we obtained are in line with low to moderate heritability
418 estimates of cognitive abilities collected on laboratory populations (reviewed in [76] see also
419 (Sauce et al, this issue) and (Sorato et al, this issue)) , and with selectively bred animals that
420 have shown large differences in, for example, numerical learning in guppies [77], oviposition
421 learning in *Drosophila* [78] and butterflies [79], or maze navigation in rats [80]. These results
422 should thus promote the investigation of individual variation in cognitive performance, ideally
423 as a first step to assessing heritability, the effect of permanent environment and experience on
424 this variation, and examining potential evolutionary consequences of this variation [6,81].

425

426 Contextual repeatability was assessed by examining performance on novel variants of the same
427 task (e.g. change of stimuli dimension) or different tasks that we considered assessed the same
428 cognitive process. Such an approach has been advocated to improve our understanding of the
429 nature of cognitive processes involved [48], (Volter et al. This issue). In line with this, our
430 estimates of contextual repeatability was moderate ($R = [0.20-0.27]$) and significant, indicating
431 that the use of different stimuli dimension, perceptual dimensions, apparatuses and tests allows
432 us to measure repeatable variation in individual cognitive performance. Of course, our
433 interpretation of R values assumes that cognitive tests are conducted in a way that minimises
434 the impact of other traits that could be repeatable as well, such as motor capacities, motivation
435 or personality traits [48].

436

437 Here, we suggest that investigators bear in mind that some possible confounds could lower
438 contextual repeatability when deploying tasks that use different stimuli or perceptual
439 dimensions. For instance, adaptive specialisations that result in differential attention to
440 particular stimuli may result in high within-individual variation in performance over contexts,
441 or in low between-individual variation in one or both contexts [82] (e.g. individuals of some
442 species may show greater variation in their performance when learning shape discrimination,
443 but relatively little variation when learning a colour discrimination task or vice versa for other
444 species, even if both tasks were under the same principle of visual-cue learning e.g. [83],[84]).
445 Using different tasks or apparatuses to examine the same putative cognitive process may also
446 lead to low contextual repeatability if the salience of stimuli differs between apparatuses. For
447 example, presenting stimuli on a touchscreen as opposed to presenting stimuli with solid objects
448 may vary the salience of stimuli [85]. Such differences may inflate within-individual variance
449 and thus decrease repeatability. Finally, while we may assume similar cognitive processes are
450 involved in a variant of the same task, we may obtain low contextual repeatability if the variants

451 require different cognitive processes. One possible solution is to conduct repeatability analyses
452 on the portion of variance likely due to a shared cognitive process by incorporating measures
453 of ‘micro-behaviours’. For example, Chow and colleagues [86] used the response latencies to
454 correct and incorrect stimuli to reflect inhibitory control, and the rate of head-switching (head-
455 turning between stimuli) to reflect attention, alongside using the number of errors in learning a
456 colour discrimination-reversal learning task on a touch screen. Assessing micro-behaviours
457 may therefore capture specific processes that are closely related to the general cognitive process
458 than more classical approaches. Accordingly, the assay of repeatability of cognitive
459 performances could then be examined by repeatedly recording a suite of micro-behavioural
460 traits as well as traditional measures of performance in the same, or variants of the same, task.

461

462 *Test order and the repeatability of cognitive performance*

463 Animals may improve their performance with increased learning/experience on the same task
464 or on a different but related task, and hence, controlling for time-related changes (i.e. the
465 number of repetitions of the same task) or task presentation order (i.e. test order) may produce
466 better estimates of repeatability [87]. However, our adjusted estimates of both temporal and
467 contextual repeatability for test order did not increase although remained significant (Table 1,
468 Figure 3). The lack of increase in the mean repeatability estimates may have indicated that
469 repetition number or task order only has a mild influence on repeatability.

470

471 Despite this, an examination of the analyses that provide estimates of temporal repeatability
472 (Table S1) suggests that there may be an optimal number of repetitions when estimating
473 individual variation in cognitive performance. Indeed, prolonged exposure to the same task may
474 reduce most, if not all, between-individual variation in performance (i.e. individuals reach a
475 plateau in performance with increased experience of the same task): high repetitions of the same
476 task (ranging from 7 to 80 repetitions) produced moderate-low repeatability (mean $R = 0.22$)
477 whereas analyses with low repetitions (ranging from 2 to 3 repetitions) produced a moderate-
478 high repeatability (mean $R = 0.42$). Increasing the number of measures of cognitive
479 performance will strengthen memory and learning on a given task, which may increase within-
480 individual variance between tests as internal and external conditions change across repetitions.
481 Likewise, memory and learning may increase within-individual variance between different
482 tasks as a result of carry-over effects. Carry-over effects on repeatability may be controlled by
483 running all tests in the same order for all subjects, and by including test number or test date for
484 a given task [87]. The effect of test order on contextual repeatability should however be treated

485 with caution, as it may be affected by the number of R estimates based on small sample size
486 studies, and may also have resulted from the fact that GLMM-based repeatability forces R to
487 be positive, in comparison to unadjusted R. Nevertheless, this confound could be used to better
488 understand how variation in the environment influences cognitive performance (i.e. plasticity)
489 when examining the evolution of cognition across different contexts.

490

491 ***Individual determinants of the repeatability of cognitive performance***

492 The addition of individual effects such as sex and age, when available, seemed to increase
493 temporal but not contextual repeatability relative to models that only included test order (Table
494 1, Figure 3). This effect on temporal repeatability may partly be because the processes that
495 underlie performance on cognitive tasks may differ between juveniles and adults. For example,
496 immature freshwater snails, *Lymnaea stagnalis*, show impaired memory for the association
497 between a light flash and the whole body withdrawal response until they reach maturity [88],
498 juvenile Australian magpies, *Cracticus tibicen*, show poorer performance on a spatial memory
499 task when tested 100 days after fledging than compared to those birds that were tested 200 and
500 300 days after fledging [15], and honeybee workers, *Apis mellifera L.*, showed impaired spatial
501 memory when tested under 16 days of age as adults than compared to their counterparts that
502 were older than 16 days [89]. Adult Eurasian harvest mice, *Micromys minutus*, also show higher
503 repeatability than juveniles on a spatial recognition task [53]. Controlling for age and
504 developmental life-stage, either experimentally (e.g. target one age group) or statistically, thus
505 seems important when assessing repeatability of cognitive performance.

506

507 Males and females may experience different selective pressures on given cognitive processes
508 that reflect different fitness consequences. Examples of such sex differences include spatial
509 orientation and reference memory in rodents [90], colour and position cues learning in chicks
510 [91], and foraging innovation in guppies [92]. Sex differences in cognitive processes may also
511 result from mating behaviours such as territory defense or mate searching, which may reduce
512 between-individual variation within the same sex. Here, we have only examined and discussed
513 a few of the individual factors that may influence the estimation of cognitive performance
514 across individuals, and thus potentially impact the estimates of repeatability. We suggest that
515 the choice of variables included in analyses of adjusted repeatability should reflect the goals of
516 the study, and include explanations of what aspects are controlled for and more importantly,
517 why [24].

518

519 *Moderators of the repeatability of cognitive performance*

520 Variation among studies used in a meta-analysis can cause heterogeneity in effect sizes that are
521 directly attributable to the experimental approach, and accounting for such variation can
522 provide insights into which factors influence the trait of interest [74]. For example, we might
523 expect that repeated measurements that are obtained after shorter time intervals may produce
524 better estimates of repeatability because the internal and external states of individuals may be
525 more similar [75]. However, our results showed that the interval between two tasks did not
526 significantly affect most estimates of temporal or contextual repeatability. Although animals
527 may form memory associations on a given test, our finding suggests that carry-over effects may
528 have minor effects on the relative extent of between vs. within-individual variation.

529

530 Among the moderators that we examined here, the type of cognitive performance measurement
531 had a strong effect on estimates of repeatability (Table 2). For contextual repeatability, the
532 lowest estimated R values are obtained for latency measures with most confidence intervals of
533 estimates overlapping with 0 (Figure S11). The very low repeatability of latency measures
534 between performance using different apparatuses may be affected by ceiling effects (e.g.
535 individuals may solve an easy task with similar latencies but show greater variation when
536 solving a more difficult problem) and floor effects (e.g. individuals may use the maximum time
537 that is given in a trial to solve a more difficult problem but show variation for an easy task)
538 [93,94]. With this in mind, the effects of internal or external variables on repeatability may be
539 minimised by using binary measures such as success-or-failure (SUC), which may ‘dilute’ the
540 effects of internal or external contextual variables. Our results indicate that certain types of
541 measurement (e.g. latency or the number of trials) used in some cognitive tasks are more
542 sensitive to internal or external contextual variables than others and thus, provide less reliable
543 measures of R. However, we suggest that moderator effects should be interpreted with caution,
544 as constraints on our sample size prevented us from controlling for other fixed effects when
545 revealing each moderator effect as well as potential interaction effects. Our approach of
546 univariate testing may thus have been more liberal than a full model approach. While our results
547 as a whole suggest that most moderators did not explain variation in the repeatability of inter-
548 individual variation in cognitive performance across studies, these factors may still be important
549 to consider when designing experiments for a particular species.

550

551 Finally, because repeatability of cognitive performance as only recently received attention, we
552 only found 6 studies reporting such estimate and had to ask around for primary dataset to

553 perform a proper cross-species meta-analysis. Such approach comes with the bias that we only
554 asked people present in the workshop “Causes and consequences of individual variation in
555 cognition” or that we knew was working on individual differences. Future meta-analysis on the
556 topic should try to incorporate a wider range of study including test-retest literature in humans
557 [33] and general intelligence studies (Dubois et al, this issue; Sauce et al, this issue).

558

559 *General conclusion and future research*

560 While we made an attempt at understanding the repeatability of cognitive performance, we
561 admit that this is an emerging field. Accordingly, this study suffers some limitations, including
562 a modest sample size (both for the number of studies included and for the number of subjects
563 provided in each study) which reduces the robustness of the conclusions regarding the effect of
564 potential moderators. Future studies may therefore benefit from the growing body of literature
565 on individual differences in cognition [81],[82],[95], this volume]. Note that other studies
566 collecting repeated measures from repetitions of a same test, or functionally-similar tests, could
567 also offer valuable datasets. In order to facilitate future meta-analyses, we suggest that authors:
568 (i) publish their datasets using the finest-grained information available (e.g. trial-by-trial instead
569 of aggregate values, such as proportion of correct choices or trials); (ii) include information on
570 potential moderators (e.g. date of test, subject’s origin) and other fixed effects (e.g. sex, age)
571 that may need to be controlled for; and (iii) include and standardise the term ‘cognitive
572 repeatability’ in their keywords.

573

574 To summarise, we report low to moderate estimates for the repeatability of cognitive
575 performance, suggesting consistent individual differences over a range of cognitive tasks and
576 taxa. Measurements of cognitive performance in a given task are thus moderately consistent for
577 individuals over time and can be studied much like other behavioral and morphological traits.
578 Furthermore, different experimental paradigms that are used to assess the same underlying
579 cognitive capacity are reasonably concordant. This suggests that different approaches can be
580 used to estimate the same underlying cognitive capacity. Together, our results suggest that
581 formally assessing individual variation in cognitive performance within populations could be a
582 useful first step in research programs on the evolutionary biology of cognition. Future avenues
583 for research may include: (1) studying the repeatability of reaction norms of cognitive
584 performance (i.e. its plasticity [96],[97] over gradients of interest, for example, deprivation
585 level or housing conditions), so as to assess the generality of the individual differences that are
586 captured by cognitive tasks across different environments and physiological states; and (2)

587 partitioning the variance among and within individuals, by making use of multiple (>4) trials
588 recorded for each individual [98]. By partitioning variance in cognitive performance at various
589 hierarchical levels (within and between individuals) we may complement approaches that
590 quantify variation at other levels (populations and species) and hence further our understanding
591 of the evolution of cognition. This approach may provide a greater understanding of the factors
592 that influence repeatability estimates, which are based on a ratio, and thus do not allow the
593 separation of variance that is due to different phenotypes (among-individual) from those due to
594 the plasticity in the response of each animal (within-individual). Separating these values could
595 provide a way to focus on the portion of variance that is expected to be heritable, and to test
596 hypotheses on the factors that affect variation within-individuals between repeated trials.

597 **References**

- 598 1. Shettleworth SJ. 2010 *Cognition, Evolution, and Behavior*. Oxford University Press.
- 599 2. van Horik J, Emery NJ. 2011 Evolution of cognition. *Wiley Interdiscip. Rev. Cogn. Sci.* **2**, 621–
600 633.
- 601 3. Van Horik JO, Clayton NS, Emery NJ. 2012 *Convergent Evolution of Cognition in Corvids, Apes*
602 *and Other Animals*.
- 603 4. MacLean EL *et al.* 2012 How does cognition evolve? Phylogenetic comparative psychology.
604 *Anim. Cogn.* **15**, 223–238.
- 605 5. Thornton A, Isden J, Madden JR. 2014 Toward wild psychometrics: linking individual cognitive
606 differences to fitness. *Behav. Ecol.* **25**, 1299–1301.
- 607 6. Cauchoix M, Chaine AS. 2016 How can we study the evolution of animal minds? *Front. Psychol.*
608 **7**, 358.
- 609 7. Dukas R, Bernays EA. 2000 Learning improves growth rate in grasshoppers. *Proc. Natl. Acad.*
610 *Sci. U. S. A.* **97**, 2637–2640.
- 611 8. Raine NE, Chittka L. 2008 The correlation of learning speed and natural foraging success in
612 bumble-bees. *Proc R Soc Lond B Biol Sci.* **275**, 803–808.
- 613 9. Pasquier G, Grüter C. 2016 Individual learning performance and exploratory activity are linked to
614 colony foraging success in a mass-recruiting ant. *Behav. Ecol.* **27**, 1702–1709.
- 615 10. Maille A, Schradin C. 2016 Survival is linked with reaction time and spatial memory in African
616 striped mice. *Biol. Lett.* **12**. (doi:10.1098/rsbl.2016.0346)
- 617 11. Kotrschal A, Buechel SD, Zala SM, Corral-Lopez A, Penn DJ, Kolm N. 2015 Brain size affects
618 female but not male survival under predation threat. *Ecol. Lett.* **18**, 646–652.
- 619 12. Keagy J, Savard J-F, Borgia G. 2009 Male satin bowerbird problem-solving ability predicts
620 mating success. *Anim. Behav.* **78**, 809–817.
- 621 13. Cole EF, Morand-Ferron J, Hinks AE, Quinn JL. 2012 Cognitive ability influences reproductive
622 life history variation in the wild. *Curr. Biol.* **22**, 1808–1812.
- 623 14. Cauchard L, Boogert NJ, Lefebvre L, Dubois F, Doligez B. 2013 Problem-solving performance is
624 correlated with reproductive success in a wild bird population. *Anim. Behav.* **85**, 19–26.
- 625 15. Ashton BJ, Ridley AR, Edwards EK, Thornton A. 2018 Cognitive performance is linked to group
626 size and affects fitness in Australian magpies. *Nature* **554**, 364–367.
- 627 16. Isden J, Panayi C, Dingle C, Madden J. 2013 Performance in cognitive and problem-solving tasks
628 in male spotted bowerbirds does not correlate with mating success. *Anim. Behav.* **86**, 829–838.
- 629 17. Dunlap AS, Stephens DW. 2016 Reliability, uncertainty, and costs in the evolution of animal
630 learning. *Curr. Opin. Behav. Sci.* **12**, 73–79.
- 631 18. Mery F. 2013 Natural variation in learning and memory. *Curr. Opin. Neurobiol.* **23**, 52–56.
- 632 19. Kawecki TJ. 2009 Evolutionary ecology of learning: insights from fruit flies. *Popul. Ecol.* **52**,
633 15–25.

- 634 20. Kotrschal A, Rogell B, Bundsen A, Svensson B, Zajitschek S, Brännström I, Immler S, Maklakov
635 AA, Kolm N. 2013 Artificial selection on relative brain size in the guppy reveals costs and
636 benefits of evolving a larger brain. *Curr. Biol.* **23**, 168–171.
- 637 21. Endler JA. 1986 *Natural Selection in the Wild*. Princeton University Press.
- 638 22. Dohm MR. 2002 Repeatability estimates do not always set an upper limit to heritability. *Funct.*
639 *Ecol.* **16**, 273–280.
- 640 23. Edwards AWF, Falconer DS. 1982 Introduction to Quantitative Genetics. *Biometrics* **38**, 1128.
- 641 24. Wilson AJ. 2018 How should we interpret estimates of individual repeatability? *Evolution Letters*
642 **2**, 4–8.
- 643 25. Dohm MR. 2002 Repeatability estimates do not always set an upper limit to heritability. *Funct.*
644 *Ecol.* **16**, 273–280.
- 645 26. Griffin AS, Guillette LM, Healy SD. 2015 Cognition and personality: an analysis of an emerging
646 field. *Trends Ecol. Evol.* **30**, 207–214.
- 647 27. Martin JGA, Réale D. 2008 Temperament, risk assessment and habituation to novelty in eastern
648 chipmunks, *Tamias striatus*. *Anim. Behav.* **75**, 309–318.
- 649 28. Bell AM, Hankison SJ, Laskowski KL. 2009 The repeatability of behaviour: a meta-analysis.
650 *Anim. Behav.* **77**, 771–783.
- 651 29. Dingemanse N, Réale D. 2005 Natural selection and animal personality. *Behaviour* **142**, 1159–
652 1184.
- 653 30. Nicolaus M, Tinbergen JM, Bouwman KM, Michler SPM, Ubels R, Both C, Kempenaers B,
654 Dingemanse NJ. 2012 Experimental evidence for adaptive personalities in a wild passerine bird.
655 *Proc. R. Soc. B* **279**, 4885–4892.
- 656 31. Dingemanse NJ, Wolf M. 2010 Recent models for adaptive personality differences: a review.
657 *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **365**, 3947–3958.
- 658 32. Dall SRX, Houston AI, McNamara JM. 2004 The behavioural ecology of personality: consistent
659 individual differences from an adaptive perspective. *Ecol. Lett.* **7**, 734–739.
- 660 33. Hughes C, Adlam A, Happé F, Jackson J, Taylor A, Caspi A. 2000 Good test—retest reliability
661 for standard and advanced false-belief tasks across a wide range of abilities. *J. Child Psychol.*
662 *Psychiatry* **41**, 483–490.
- 663 34. Guenther A, Brust V. 2017 Individual consistency in multiple cognitive performance:
664 behavioural versus cognitive syndromes. *Anim. Behav.* **130**, 119–131.
- 665 35. Brust V, Guenther A. 2017 Stability of the guinea pigs personality - cognition - linkage over
666 time. *Behav. Processes* **134**, 4–11.
- 667 36. Gibelli J, Dubois F. 2016 Does personality affect the ability of individuals to track and respond to
668 changing conditions? *Behav. Ecol.* **28**, 101–107.
- 669 37. Ashton BJ, Ridley AR, Edwards EK, Thornton A. 2018 Cognitive performance is linked to group
670 size and affects fitness in Australian magpies. *Nature* **554**, 364–367.
- 671 38. Tello-Ramos MC, Branch CL, Pitera AM, Kozlovsky DY, Bridge ES, Pravosudov VV. 2018
672 Memory in wild mountain chickadees from different elevations: comparing first-year birds with

- 673 older survivors. *Anim. Behav.* **137**, 149–160.
- 674 39. Chittka L, Dyer AG, Bock F, Dornhaus A. 2003 Psychophysics: bees trade off foraging speed for
675 accuracy. *Nature* **424**, 388.
- 676 40. Dalesman S, Rendle A, Dall SRX. 2015 Habitat stability, predation risk and ‘memory
677 syndromes’. *Sci. Rep.* **5**. (doi:10.1038/srep10538)
- 678 41. Morand-Ferron J, Cole EF, Quinn JL. 2016 Studying the evolutionary ecology of cognition in the
679 wild: a review of practical and conceptual challenges. *Biol. Rev. Camb. Philos. Soc.* **91**, 367–389.
- 680 42. Rowe C, Healy SD. 2014 Measuring variation in cognition. *Behav. Ecol.* **25**, 1287–1292.
- 681 43. van Horik JO, Langley EJG, Whiteside MA, Laker PR, Beardsworth CE, Madden JR. 2018 Do
682 detour tasks provide accurate assays of inhibitory control? *Proc. Biol. Sci.* **285**.
683 (doi:10.1098/rspb.2018.0150)
- 684 44. Dingemanse NJ, Dochtermann NA. 2013 Quantifying individual variation in behaviour: mixed-
685 effect modelling approaches. *J. Anim. Ecol.* **82**, 39–54.
- 686 45. Wilson AJ. 2018 How should we interpret estimates of individual repeatability? *Evolution Letters*
687 **2**, 4–8.
- 688 46. Nakagawa S, Schielzeth H. 2010 Repeatability for Gaussian and non-Gaussian data: a practical
689 guide for biologists. *Biol. Rev. Camb. Philos. Soc.* **85**, 935–956.
- 690 47. Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group. 2009 Preferred Reporting
691 Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Med.* **6**,
692 e1000097.
- 693 48. Griffin AS, Guillette LM, Healy SD. 2015 Cognition and personality: an analysis of an emerging
694 field. *Trends Ecol. Evol.* **30**, 207–214.
- 695 49. Hughes C, Adlam A, Happé F, Jackson J, Taylor A, Caspi A. 2000 Good test—retest reliability
696 for standard and advanced false-belief tasks across a wide range of abilities. *J. Child Psychol.*
697 *Psychiatry* **41**, 483–490.
- 698 50. Duckworth AL, Kern ML. 2011 A meta-analysis of the convergent validity of self-control
699 measures. *J. Res. Pers.* **45**, 259–268.
- 700 51. Rodríguez RL, Gloudeman MD. 2011 Estimating the repeatability of memories of captured prey
701 formed by *Frontinella communis* spiders (Araneae: Linyphiidae). *Anim. Cogn.* **14**, 675–682.
- 702 52. Guenther A, Brust V. 2017 Individual consistency in multiple cognitive performance:
703 behavioural versus cognitive syndromes. *Anim. Behav.* **130**, 119–131.
- 704 53. Schuster AC, Carl T, Foerster K. 2017 Repeatability and consistency of individual behaviour in
705 juvenile and adult Eurasian harvest mice. *Naturwissenschaften* **104**, 10.
- 706 54. Schuster AC, Zimmermann U, Hauer C, Foerster K. 2017 A behavioural syndrome, but less
707 evidence for a relationship with cognitive traits in a spatial orientation context. *Front. Zool.* **14**,
708 19.
- 709 55. Shaw RC. 2017 Testing cognition in the wild: factors affecting performance and individual
710 consistency in two measures of avian cognition. *Behav. Processes* **134**, 31–36.
- 711 56. Cole EF, Cram DL, Quinn JL. 2011 Individual variation in spontaneous problem-solving

- 712 performance among wild great tits. *Anim. Behav.* **81**, 491–498.
- 713 57. Koricheva J, Gurevitch J, Mengersen K. 2013 *Handbook of Meta-analysis in Ecology and*
714 *Evolution*. Princeton University Press.
- 715 58. R Development Core Team. 2017 *R: A Language and Environment for Statistical Computing*.
- 716 59. Stoffel MA, Nakagawa S, Schielzeth H. 2017 rptR: repeatability estimation and variance
717 decomposition by generalized linear mixed-effects models. *Methods Ecol. Evol.* **8**, 1639–1644.
- 718 60. Lessells CM, Boag PT. 1987 Unrepeatable Repeatabilities: A Common Mistake. *Auk* **104**, 116–
719 121.
- 720 61. Holtmann B, Santos ESA, Lara CE, Nakagawa S. 2017 Personality-matching habitat choice,
721 rather than behavioural plasticity, is a likely driver of a phenotype-environment covariance. *Proc*
722 *R Soc Lond B Biol Sci.* **284**, 20170943.
- 723 62. Holtmann B, Lagisz M, Nakagawa S. 2016 Metabolic rates, and not hormone levels, are a likely
724 mediator of between-individual differences in behaviour: a meta-analysis. *Funct. Ecol.* **31**, 685–
725 696.
- 726 63. Wolak ME, Fairbairn DJ, Paulsen YR. 2011 Guidelines for estimating repeatability. *Methods*
727 *Ecol. Evol.* **3**, 129–137.
- 728 64. Viechtbauer W. 2010 Conducting Meta-Analyses in R with the metafor Package. *J. Stat. Softw.*
729 **36**. (doi:10.18637/jss.v036.i03)
- 730 65. Hinchliff CE *et al.* 2015 Synthesis of phylogeny and taxonomy into a comprehensive tree of life.
731 *Proc. Natl. Acad. Sci. U. S. A.* **112**, 12764–12769.
- 732 66. Michonneau F, Brown JW, Winter D. 2016 rotl, an R package to interact with the Open Tree of
733 Life data. (doi:10.7287/peerj.preprints.1471)
- 734 67. Nakagawa S, Schielzeth H. 2012 The mean strikes back: mean–variance relationships and
735 heteroscedasticity. *Trends Ecol. Evol.* **27**, 474–475.
- 736 68. Benjamini Y, Hochberg Y. 1995 Controlling the false discovery rate: a practical and powerful
737 approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **57**, 289–300.
- 738 69. Higgins JPT, Thompson SG. 2002 Quantifying heterogeneity in a meta-analysis. *Stat. Med.* **21**,
739 1539–1558.
- 740 70. Nakagawa S, Santos ESA. 2012 Methodological issues and advances in biological meta-analysis.
741 *Evol. Ecol.* **26**, 1253–1274.
- 742 71. Egger M, Davey Smith G, Schneider M, Minder C. 1997 Bias in meta-analysis detected by a
743 simple, graphical test. *BMJ* **315**, 629–634.
- 744 72. Nakagawa S, Santos ESA. 2012 Methodological issues and advances in biological meta-analysis.
745 *Evol. Ecol.* **26**, 1253–1274.
- 746 73. Moreno SG, Sutton AJ, Ades AE, Stanley TD, Abrams KR, Peters JL, Cooper NJ. 2009
747 Assessment of regression-based methods to adjust for publication bias through a comprehensive
748 simulation study. *BMC Med. Res. Methodol.* **9**, 2.
- 749 74. Nakagawa S, Noble DWA, Senior AM, Lagisz M. 2017 Meta-evaluation of meta-analysis: ten
750 appraisal questions for biologists. *BMC Biol.* **15**, 18.

- 751 75. Bell AM, Hankison SJ, Laskowski KL. 2009 The repeatability of behaviour: a meta-analysis.
752 *Anim. Behav.* **77**, 771–783.
- 753 76. Croston R, Branch CL, Kozlovsky DY, Dukas R, Pravosudov VV. 2015 Heritability and the
754 evolution of cognitive traits. *Behav. Ecol.* **26**, 1447–1459.
- 755 77. Kotrschal A, Rogell B, Bundsen A, Svensson B, Zajitschek S, Brännström I, Immler S, Maklakov
756 AA, Kolm N. 2013 Artificial selection on relative brain size in the guppy reveals costs and
757 benefits of evolving a larger brain. *Curr. Biol.* **23**, 168–171.
- 758 78. Burger JMS, Kolss M, Pont J, Kawecki TJ. 2008 Learning ability and longevity: a symmetrical
759 evolutionary trade-off in *Drosophila*. *Evolution* **62**, 1294–1304.
- 760 79. Snell-Rood EC, Davidowitz G, Papaj DR. 2011 Reproductive tradeoffs of learning in a butterfly.
761 *Behav. Ecol.* **22**, 291–302.
- 762 80. Tryon RC. 1940 Studies in individual differences in maze ability. VII. The specific components
763 of maze ability, and a general theory of psychological components. *J. Comp. Psychol.* **30**, 283–
764 335.
- 765 81. Thornton A, Lukas D. 2012 Individual variation in cognitive performance: developmental and
766 evolutionary perspectives. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **367**, 2773–2783.
- 767 82. Rowe C, Healy SD. 2014 Measuring variation in cognition. *Behav. Ecol.* **25**, 1287–1292.
- 768 83. Wäckers FL, Lewis WJ. 1999 A comparison of color-, shape- and pattern-learning by the
769 hymenopteran parasitoid *Microplitis croceipes*. *J. Comp. Physiol. A* **184**, 387–393.
- 770 84. Aronsson M, Gamberale-Stille G. 2008 Domestic chicks primarily attend to colour, not pattern,
771 when learning an aposematic coloration. *Anim. Behav.* **75**, 417–423.
- 772 85. O’Hara M, Huber L, Gajdon GK. 2015 The advantage of objects over images in discrimination
773 and reversal learning by kea, *Nestor notabilis*. *Anim. Behav.* **101**, 51–60.
- 774 86. Chow PKY, Leaver LA, Wang M, Lea SEG. 2017 Touch screen assays of behavioural flexibility
775 and error characteristics in Eastern grey squirrels (*Sciurus carolinensis*). *Anim. Cogn.* **20**, 459–
776 471.
- 777 87. Biro PA, Stamps JA. 2015 Using repeatability to study physiological and behavioural traits:
778 ignore time-related change at your peril. *Anim. Behav.* **105**, 223–230.
- 779 88. Ono M, Kawai R, Horikoshi T, Yasuoka T, Sakakibara M. 2002 Associative learning acquisition
780 and retention depends on developmental stage in *Lymnaea stagnalis*. *Neurobiol. Learn. Mem.* **78**,
781 53–64.
- 782 89. Ushitani T, Perry CJ, Cheng K, Barron AB. 2016 Accelerated behavioural development changes
783 fine-scale search behaviour and spatial memory in honey bees (*Apis mellifera* L.). *J. Exp. Biol.*
784 **219**, 412–418.
- 785 90. Jonasson Z. 2005 Meta-analysis of sex differences in rodent models of learning and memory: a
786 review of behavioral and biological data. *Neurosci. Biobehav. Rev.* **28**, 811–825.
- 787 91. Vallortigara G. 1996 Learning of colour and position cues in domestic chicks: Males are better at
788 position, females at colour. *Behav. Processes* **36**, 289–296.
- 789 92. Laland KN, Reader SM. 1999 Foraging innovation in the guppy. *Anim. Behav.* **57**, 331–340.

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

805

806 **Figure and table captions**

807 Figure 1: Temporal repeatability R (unadjusted) and 95% bootstrapped confidence intervals for
808 each dataset. First author, species name, cognitive task and cognitive measurement are indicated
809 on the y-axis. Cognitive performance measurement was the quantification of a cognitive
810 process using: accuracy, e.g. proportion correct (ACC); the number of trials to reach a learning
811 criterion (TTC); success-or-failure binary outcome (SUC); latency (LAT); normalised
812 performance scores (NOR); the number of correct trials or errors over a fixed number of trials
813 (NBT). Cognitive task type included: mechanical problem solving (PS); discriminative learning
814 (DL); reversal learning (RL); inhibition (IN); memory (ME); use of human cue (HC); external
815 attention (EA); internal attention (IA); learning (LE); Physical cognition (PC) that include
816 visual exclusion performance; auditory exclusion performance and object permanence; social
817 learning (SL), spatial orientation learning (SOL), spatial recognition (SR) and lexical fluency
818 (LF).

819

820 Figure 2: Contextual repeatability R (unadjusted) and 95% bootstrapped confidence intervals
821 for each dataset. First author, species name, cognitive task and cognitive measurement are
822 indicated on the y-axis. Cognitive performance measurement was the quantification of a
823 cognitive process using: accuracy, e.g. proportion correct (ACC); the number of trials to reach
824 a learning criterion (TTC); success-or-failure binary outcome (SUC); latency (LAT);
825 normalised performance scores (NOR); the number of correct trials or errors over a fixed

826 number of trials (NBT). Cognitive task type included: mechanical problem solving (PS);
827 discriminative learning (DL); reversal learning (RL); inhibition (IN); memory (ME); use of
828 human cue (HC); external attention (EA); internal attention (IA); learning (LE); Physical
829 cognition (PC) that include visual exclusion performance; auditory exclusion performance and
830 object permanence; social learning (SL), spatial orientation learning (SOL), spatial recognition
831 (SR) and lexical fluency (LF).

832

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

838

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

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

845