The repeatability of cognitive performance: a meta-analysis

Maxime Cauchoix, P K y Chow, J O van Horik, C M Atance, E. Barbeau, Gladys Barragan-Jason, P. Bize, A Boussard, S D Buechel, Amélie Cabirol, et al.

PHILOSOPHICAL TRANSACTIONS B

rstb.royalsocietypublishing.org

Research

Cite this article: Cauchoix M *et al*. 2018 The repeatability of cognitive performance: a meta-analysis. *Phil. Trans. R. Soc. B* **373**: 20170281.
http://dx.doi.org/10.1098/rstb.2017.0281

Accepted: 28 June 2018

One contribution of 15 to a theme issue 'Causes and consequences of individual differences in cognitive abilities'.

Subject Areas:
behaviour, cognition, evolution

Keywords:
cognitive repeatability, evolutionary biology of cognition, individual differences, learning, memory, attention

Author for correspondence:
M. Cauchoix
e-mail: mcauchoixxx@gmail.com

†Shared first authorship listed alphabetically.
‡Shared senior authorship listed alphabetically.

Electronic supplementary material is available online at https://doi.org/10.6084/m9.figshare.c.4153862.

THE ROYAL SOCIETY PUBLISHING

The repeatability of cognitive performance: a meta-analysis

M. Cauchoix[1,2,†], P. K. Y. Chow[3,4,†], J. O. van Horik[3,†], C. M. Atance[5], E. J. Barbeau[6], G. Barragan-Jason[2], P. Bize[7], A. Boussard[8], S. D. Buechel[8], A. Cabirol[9], L. Cauchard[10], N. Claidière[11], S. Dalesman[12], J. M. Devaud[9], M. Didic[13], B. Doligez[14], J. Fagot[11], C. Fichtel[15,16,17], J. Henke-von der Malsburg[15,16,17], E. Hermer[16], L. Huber[17], F. Huebner[15,16,17], P. M. Kappeler[15,16,17], S. Klein[9], J. Langbein[20], E. J. G. Langley[3], S. E. G. Lea[3], M. Lihoreau[9], H. Lovlie[21], L. D. Matzel[22], S. Nakagawa[23], C. Nawroth[20], S. Oesterwind[24], B. Sauce[22], E. A. Smith[25], E. Sorato[21], S. Tebbich[26], L. J. Wallis[19,27], M. A. Whiteside[3], A. Wilkinson[25], A. S. Chaine[1,2,‡] and J. Morand-Ferron[18,‡]

[1]Station d'Ecologie Théorique et Expérimentale du CNRS UMR5321, Evolutionary Ecology Group, 2 route du CNRS, 09200 Moulis, France
[2]Institute for Advanced Study in Toulouse, 21 allée de Brienne, 31015 Toulouse, France
[3]Centre for Research in Animal Behaviour, Psychology, University of Exeter, Exeter, UK
[4]Graduate School of Environmental Science, Division of Biospohere Science, Hokkaido University, Sapporo, Hokkaido, Japan
[5]School of Psychology, University of Ottawa, Ottawa, Canada
[6]Centre de recherche Cerveau et Cognition, UPS-CNRS, UMR5549, Toulouse, France
[7]Institute of Biological and Environmental Sciences, University of Aberdeen, Aberdeen, UK
[8]Department of Zoology/Ethology, Stockholm University, Svante Arrheniusväg 18B, 10691 Stockholm, Sweden
[9]Research Center on Animal Cognition (CRCA), Center for Integrative Biology (CBI), CNRS, University Paul Sabatier, Toulouse, France
[10]Département de Sciences Biologiques, Université de Montréal, Montreal, Quebec, Canada
[11]LPC, Aix Marseille University, CNRS, Marseille, France
[12]Institute of Biological, Environmental and Rural Sciences, Aberystwyth University, Aberystwyth, UK
[13]AP-HM Timone & Institut de Neurosciences des Systèmes, Marseille, France
[14]Department of Biometry and Evolutionary Biology, CNRS UMR 5558, Université Lyon 1, Université de Lyon, Villeurbanne, France
[15]Behavioural Ecology and Sociobiology Unit, German Primate Centre, Leibniz Institute for Primatology, Kellnerweg 4, 37077 Göttingen, Germany
[16]Department of Sociobiology/Anthropology, Johann-Friedrich-Blumenbach Institute for Zoology and Anthropology, University of Göttingen, Kellnerweg 6, 37077 Göttingen, Germany
[17]Leibniz Science Campus 'Primate Cognition', Göttingen, Germany
[18]Department of Biology, University of Ottawa, Ottawa, Canada
[19]Clever Dog Lab, Messerli Research Institute, University of Veterinary Medicine Vienna, Medical University of Vienna, University of Vienna, Vienna, Austria
[20]Institute of Behavioural Physiology, Leibniz Institute for Farm Animal Biology, Dummerstorf, Germany
[21]IFM Biology, Linköping University, 58183 Linköping, Sweden
[22]Department of Psychology, Rutgers University, Piscataway, NJ, USA
[23]Evolution & Ecology Research Centre and School of Biological, Earth & Environmental Sciences, University of New South Wales, Sydney, New South Wales 2052, Australia
[24]Faculty of Agricultural and Environmental Sciences, University of Rostock, Rostock, Germany
[25]School of Life Sciences, University of Lincoln, Lincoln, UK
[26]Department of Behavioural Biology, University of Vienna, Vienna, Austria
[27]Department of Ethology, Eötvös Loránd University, Budapest, Hungary

MC, 0000-0002-8233-6311; PKYC, 0000-0002-8208-592X; JOvH, 0000-0002-8319-911X; EJB, 0000-0003-0836-3538; PB, 0000-0002-6759-4371; SDB, 0000-0002-2385-2973; SD, 0000-0001-8548-3096; BD, 0000-0003-3015-5022; JH-vdM, 0000-0001-6055-8506; LH, 0000-0002-0217-136X; FH, 0000-0001-9583-1680; PMK, 0000-0002-4801-487X; SK, 0000-0001-8815-6250; JL, 0000-0002-1170-5431; EJGL, 0000-0001-8980-8206; ML, 0000-0002-2463-2040; HL, 0000-0003-4352-6275; LDM, 0000-0003-0462-7188; SN, 0000-0002-7765-5182; CN, 0000-0003-4582-4057; BS, 0000-0002-9544-0150; ES, 0000-0002-5223-4496; AW, 0000-0002-4500-0181; ASC, 0000-0003-3346-551X

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

This article is part of the theme issue 'Causes and consequences of individual differences in cognitive abilities'.

1. Introduction

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

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

Assessing the repeatability of behavioural or cognitive traits is, however, challenging, because the context of measurement can influence the behaviour of animals, and thus the value recorded. Contextual variation can come from the internal state of the organism (e.g. hunger, circadian cycle, recent interactions, stress) and/or the external environment, which may differ between trials [22]. Moreover, behavioural and cognitive measures may suffer further variation between measures as experience with one type of measure or test can influence subsequent measures via processes such as learning and memory [25]. While this issue has been recognized and discussed in recent research on animal personality [26], it may be particularly relevant when assaying the repeatability of cognitive traits. Consequently, we might therefore expect higher within-individual variation in behavioural or cognitive measures compared with morphological or physiological measures, owing to greater differences in the context (internal and/or external) of repeated sampling.

Research on animal personality has provided a broad understanding that individual differences in behaviour are repeatable across time and contexts (average $R = 0.37$, $R = 0.29$, $R = 0.41$: see [27–29] respectively), hence revealing an important platform for selection to act on [30–33]. Yet, relatively little is known about the stability of inter-individual variation in cognitive traits, such as those associated with learning and memory [25]. Some examples of repeatability estimates suggest that children show good test–retest reliability on false-belief tasks used to assess theory-of-mind [25,34]. Consistent individual differences in performance on cognitive tasks have also been documented in a few non-human animals, such as guinea pigs, *Cavia aperea f. porcellus* [35,36], zebra finch, *Taenopigya guttata* [37], Australian magpies, *Gymnorhina tibicen* [15], mountain chickadees, *Poecile gambeli* [38], bumblebees, *Bombus terrestris* [39], and snails, *Lymnaea stagnalis* [40]. While the paucity of repeatability measures of cognitive performance may stem from the recency of interest in the evolutionary ecology of cognitive traits [41,42], it may also suggest that it is difficult to accurately capture repeatable measures of cognitive ability [43]. Further investigation into the consistency of individual differences in cognition and how internal and external factors may influence repeatability estimates of these measures is therefore warranted.

Recent advances in analytical techniques, such as the use of mixed-effect models, have facilitated the assessment of repeatability of behavioural traits, by accounting for the potential confounding effects of both internal and external contextual variations [24,44]. Such approaches can help provide more accurate estimates of repeatability of cognitive traits and could provide new insights to the influence of internal and external factors on cognitive performance. For example, we can now explicitly address the effect of time, or an individual's condition, on the repeatability of traits of interest such as learning performance. Likewise, we can examine the effect of external factors, for example by modelling the environment (e.g. group size at testing) or the type of

test employed (e.g. spatial versus colour cues in associative learning). Adopting these methods (i.e. adjusted repeatability [45]) could therefore facilitate studies that generate repeatability estimates of cognitive performance and provide greater clarity concerning the sources of variation in measures of cognition in this rapidly expanding field.

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

2. Material and methods

(a) Data collection

We followed the preferred reporting items for systematic reviews and meta-analyses (PRISMA) approach for the collation of the datasets used in the current study [46]. We first collected published repeatability estimates of cognitive performance (electronic supplementary material, figure S1). We did not include studies reporting inter-class correlations (Pearson or Spearman) between cognitive performances on tasks measuring different cognitive abilities (i.e. general intelligence or '*g*') as we considered these outside the scope of this meta-analysis. Although we acknowledge that results from the literature on test–retest [25,34] or convergent validity [47] in psychology would be relevant to compare with the present study, we also considered them beyond the scope of this paper as their inclusion would have led to a heavy bias towards studies on humans. We only found six publications reporting repeatability values for cognitive performance (*R*) in six different species: one arachnid [48], two mammals [35,49,50] and three birds [15,51,52], with a sample size ranging from 15 to 347 (mean: 54.7, median: 33) and number of repeated tests varying from 2 to 4 (mean: 2.5, median: 2).

To complement our dataset from published studies, we used an 'individual-patient-data' meta-analysis approach commonly used in medical research [53] in which effect sizes are extracted using the same analysis on primary data [53]. We invited participants from a workshop on the 'Causes and consequences of individual variation in cognitive ability' (36 people), as well as 25 colleagues working on individual differences in cognition, to contribute primary datasets of repeated measurements of cognitive performance. From this approach, we assembled 38 primary datasets from unpublished (nine datasets: six were fully unpublished, while three had similar methods published from the same laboratory group) or published sources (29 datasets: including repeated measures of cognitive performance but that did not report *R* values) that we could use to compute repeatability using consistent analytical methods (electronic supplementary material, figure S1, see shared repository link). These datasets comprised 20 different species of mammals (humans included), insects, molluscs, reptiles and birds (electronic supplementary material, tables S1 and S2). Details about subjects, experimental context and cognitive tasks for each dataset can be found in electronic supplementary material, methods (https://doi.org/10.6084/m9.figshare.6431549.v1).

Each dataset included 4–375 individuals (mean: 46.6, median: 29) that performed 2–80 (mean: 7.9, median: 2) repetitions of tests targeting the same cognitive process, by conducting either the same task presented at different points in time (*temporal repeatability*, see electronic supplementary material, table S1), or different tasks aimed at assessing the same underlying cognitive process but using a different protocol (*contextual repeatability*, see electronic supplementary material, table S2). Tasks considered to assess contextual repeatability differed by stimulus dimension (e.g. spatial versus colour reversal learning in Cauchoix great tit dataset), sensory modality (e.g. visual versus olfactory discrimination in Henke von der Malsburg microcebus dataset), or change in experimental apparatus (e.g. colour discrimination on touch screen and on solid objects in Chow squirrel laboratory dataset) or could be a different task designed to measure the same cognitive process (i.e. Mouse Stroop Test and the Dual Radial Arm Maze to measure external attention in Matzel attention mice dataset).

(b) Repeatability analysis for primary data

All analyses were performed in the R environment for statistical computing v. 3.3.3 [54]. We performed the same repeatability analysis for all primary data provided by co-authors: (1) We first transformed cognitive variables to meet assumptions of normality; (2) To assess whether time-related changes (i.e. the number of repetitions of the same task or test order of different tasks), and/or an individual's sex and age (hereafter, individual determinants) played a role in repeatability of cognitive performances, we then computed three types of repeatability values with a mixed-effects model approach using the appropriate link function in the 'rptR' package [55]. Specifically, we calculated unadjusted repeatability (*R*), repeatability adjusted for test order (R_n), and repeatability adjusted for test order and individual determinants (R_{ni}) for *temporal* and *contextual* repeatability separately; (3) For cases with unadjusted *R* close to 0 (less than 0.005), we computed the *R* estimate using a least-squares ANOVA approach as advised in [29,56,57] using the 'ICC' package [58]; and (4) We removed *R* estimates from further analyses when residuals were not normal or overdispersed (for Poisson distribution) and for data that could not be transformed to achieve normality (see the electronic supplementary material general methods for more details; excluded *R* estimates are presented in table S3).

(c) Meta-analysis and meta-regression

We collated the 178 *R* values computed from primary data with the 35 *R* values from published studies to obtain a total of 213 estimates of cognitive repeatability. We did not recompute repeatability de novo for published studies that provide repeatability values as the statistics used in these papers are the same or similar to those used here for primary data (e.g. mixed-model approach with or without 'rptR' package). We then used a meta-analytic approach to examine average *R* estimates across species of cognitive performance. This approach allowed us to: (1) take into account sample size and number of repeated measures associated with each *R* value in the estimation of average cognitive repeatability; (2) control for repeated samples (i.e. avoid pseudoreplication) of the same species (taxonomic bias), the same laboratory group (i.e. same senior author; observer bias) or

first author	species	task	measurement
Barbeau	*Homo sapiens*	ME	NBT
Matzel	*Mus musculus*	SOL	NOR
Barbeau	*Homo sapiens*	ME	NBT
Barbeau	*Homo sapiens*	ME	NBT
Barbeau	*Homo sapiens*	ME	NOR
Barbeau	*Homo sapiens*	LF	NBT
Barbeau	*Homo sapiens*	ME	NBT
Cauchard	*Parus major*	PS	ACC
Barbeau	*Homo sapiens*	ME	NBT
Cole	*Parus major*	PS	SUC
Langbein	*Capra hircus*	RL	TTC
Barbeau	*Homo sapiens*	LF	NBT
Barbeau	*Homo sapiens*	ME	NBT
Barbeau	*Homo sapiens*	ME	NOR
Barbeau	*Homo sapiens*	ME	ACC
Cole	*Parus major*	PS	SUC
Cole	*Parus major*	PS	SUC
Cole	*Parus major*	PS	SUC
Cole	*Parus major*	PS	SUC
Matzel	*Mus musculus*	LE	NOR
Bize	*Ficedula albicollis*	PS	LAT
Cole	*Parus major*	PS	SUC
Nawroth	*Capra hircus*	PC	ACC
Claidière	*Papio papio*	ME	ACC
Matzel	*Mus musculus*	SOL	NOR
Schuster	*Micromys minutus*	SR	LAT
Schuster	*Micromys minutus*	SR	LAT
Cauchard	*Parus major*	PS	LAT
Schuster	*Micromys minutus*	SOL	TTC
van Horik	*Diopsittaca nobilis*	RL	TTC
Matzel	*Mus musculus*	DL	NOR
Matzel	*Mus musculus*	LE	NOR
Cauchoix	*Parus major*	RL	TTC
Cauchoix	*Parus major*	RL	ACC
Cauchoix	*Parus major*	RL	ACC
Dalesman	*Lymnaea stagnalis*	ME	NOR
Barbeau	*Homo sapiens*	ME	NBT
van Horik	*Pionites melanocephala*	RL	TTC
Chow	*Sciurus carolinensis*	PS	LAT
Wilkinson	*Chelonoidis carbonaria*	RL	TTC

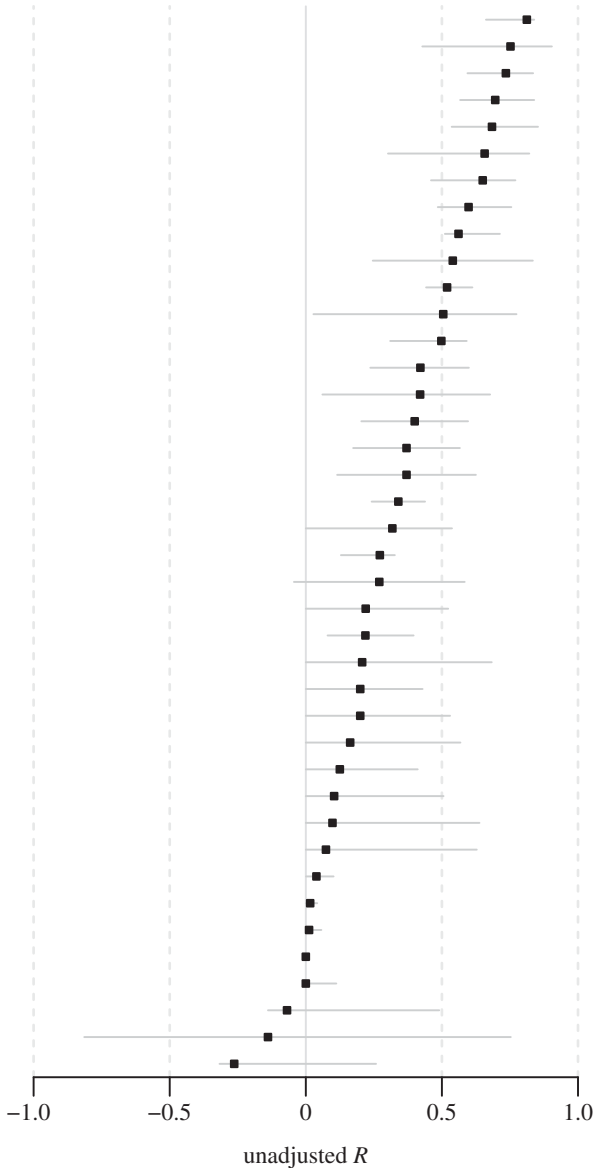

Figure 1. Temporal repeatability *R* (unadjusted) and 95% bootstrapped confidence intervals for each dataset. *Y*-axis provides information about first author, species name, the type of cognitive task and the type of cognitive performance measurement. Cognitive performance measurement was the quantification of a cognitive process using accuracy such as proportion correct (ACC); the number of trials to reach a learning criterion (TTC); success-or-failure binary outcome (SUC); latency (LAT); normalized performance scores (NOR); the number of correct trials or errors over a fixed number of trials (NBT). The types of cognitive task include: mechanical problem solving (PS); discriminative learning (DL); reversal learning (RL); memory (ME); learning (LE); physical cognition (PC), which includes visual exclusion performance, auditory exclusion performance and object permanence; spatial orientation learning (SOL); spatial recognition (SR); and lexical fluency (LF).

the same experiment (measurement bias) by including these factors as random effects; and (3) ask whether other specific factors (fixed effects called 'moderators' in meta-analysis, see below) could explain the variation in repeatability of cognitive tests.

For each of the six types of *R* analysis (i.e. unadjusted temporal *R*, adjusted temporal *R* for test order, adjusted temporal *R* for test order and individual determinants, unadjusted contextual *R*, adjusted contextual *R* for test order, adjusted contextual *R* for test order and individual determinants), we performed three different multilevel meta-analyses by fitting linear mixed models (LMMs) using the 'metafor' package [59]: (1) a standard meta-analytic model (intercept-only model) to estimate the overall mean effect size, (2) seven univariate (multilevel) meta-regression models to independently test the significance of each moderator. For each model, we used standardized (Fisher's Z transformed) *R* values as the response variable. Finally, we conducted (3) a type of Egger's regression to test for selection bias.

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

In meta-regression models, we accounted for variance in repeatability of cognitive performance by adding both fixed and random effects. We accounted for variation in repeatability related to fixed effects by including moderators. We considered seven moderators (detailed in the electronic supplementary material, general methods and figures 1 and 2): type of cognitive performance measurement (e.g. success or failure, latency, the number of trials before reaching a learning criterion); type of cognitive task (e.g. reversal learning, discrimination learning); median delay between tests; experimental context (conducted in the wild or in captivity); the origin of subjects (wild or hand-raised), taxonomic class and publication status (whether the *R* value was published or unpublished). We also took into account non-independence of data by including random effects, including species (multiple

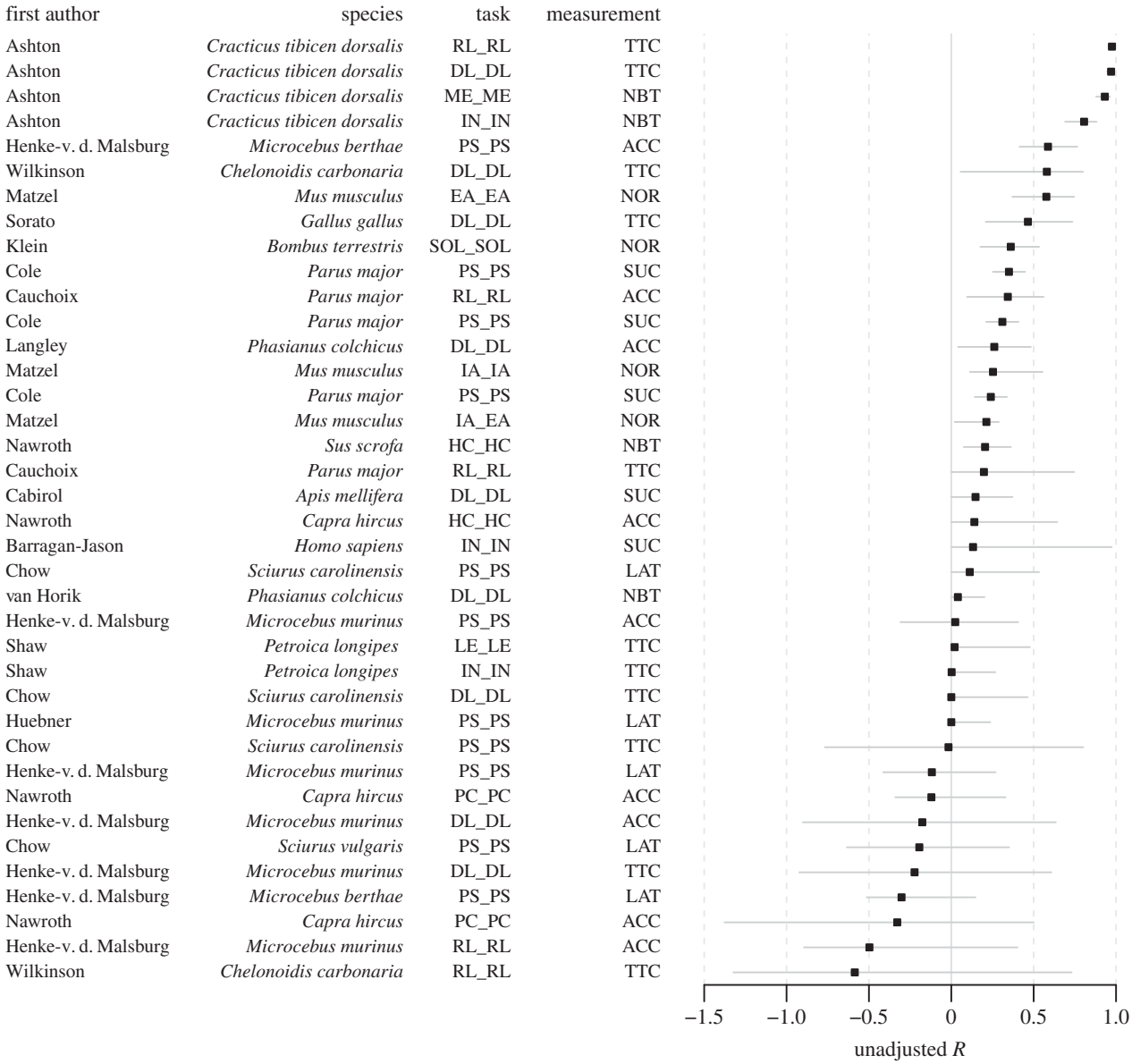

Figure 2. Contextual repeatability *R* (unadjusted) and 95% bootstrapped confidence intervals for each dataset. *Y*-axis presents first author, species name, the type of cognitive task and the type of cognitive performance measurement. Cognitive measurement is used to quantify a cognitive process using accuracy such as proportion correct (ACC); the number of trials to reach a learning criterion (TTC); success-or-failure binary outcome (SUC); latency (LAT); normalized performance scores (NOR); the number of correct trials or errors over a fixed number of trials (NBT). The types of cognitive task include: mechanical problem solving (PS); discriminative learning (DL); reversal learning (RL); inhibition (IN); memory (ME); use of human cue (HC); external attention (EA); internal attention (IA); learning (LE); physical cognition (PC) that includes visual exclusion performance, auditory exclusion performance and object permanence; and spatial orientation learning (SOL).

datasets from the same species), laboratory groups (experiments conducted by the same principal investigator) and experiments (experiments on the same subjects; see the electronic supplementary material, general methods for more details).

We controlled for the possibility that phylogenetic history influences the repeatability of cognitive abilities (i.e. closely related species may be more likely to show similar estimates of cognitive repeatability) by using a covariance matrix based on an order-level phylogenetic tree (using Open Tree of Life [60] and 'rotl' R package [61]) but only in the intercept-only model as meta-regression models failed to converge with this additional information. We ran the intercept-only meta-analysis with and without controlling for the effect of phylogeny and found that phylogenetic relationships had negligible effects on average repeatability of cognitive abilities (electronic supplementary material, table S5), justifying its exclusion in subsequent meta-regression models.

For meta-regressions, we report conditional $R2$ (*sensu* [62]), which quantifies the proportion of variance explained by fixed (moderators) and random effects along with *p*-values from omnibus tests [59], which test the significance of multiple moderator effects. When omnibus tests were significant ($p < 0.05$), we ran the same meta-regression model without the intercept to compute and plot beta coefficients associated with each level of the moderator (electronic supplementary material, figures S10 and S11) and performed multiple pairwise comparisons to estimate statistical differences between all combinations of moderator levels. We corrected for multiple comparisons using a false discovery rate adjustment of *p*-values [63].

We assessed the extent of variation among effect sizes in each meta-analytic model (intercept only) by calculating heterogeneities (I^2). Along with the overall heterogeneity (I^2_{total}), which represents between-study variance divided by the total variance [64], we also provide estimates of heterogeneity for each random factor (species, laboratory and experiment) following [65]. I^2 values of 25, 50 and 75% are generally considered to be low, moderate and high levels of heterogeneity, respectively [64].

Finally, we statistically tested for selection bias in the dataset by conducting a type of Egger's regression [66]. Given that effect

sizes were not always independent from each other (i.e. some came from the same study), we employed a mixed-model version of Egger's regression using the full models (seven moderators as fixed effects) with the sampling s.e. of each effect size as a moderator [65,67]; a regression slope of the s.e. significantly different from zero indicates selection bias [66]. Such a significant effect usually indicates that large effect sizes with large sampling variance (small sample size) are more prevalent than expected, potentially overestimating the overall effect size (i.e. R).

3. Results

(a) Dataset summary

Repeatability estimates computed from primary data are presented together with published R values in electronic supplementary material, table S1 for temporal repeatability and electronic supplementary material, table S2 for contextual repeatability. For temporal repeatability, we used 22 studies on 15 species in which 4 to 375 (mean: 56.3, median: 40) individuals performed a median of 2, 95%CI [1.91, 2.11] repeated tests, leading to a total of 106 repeatability analyses (40 R; 40 R_n; and 26 R_{ni}). For contextual repeatability, we used 27 studies on 20 species in which 4 to 297 (mean: 41, median: 24) individuals performed a median of 2, 95%CI [1.80, 2.15] repeated tests, leading to a total of 107 repeatability analyses (38 R; 32 R_n; and 37 R_{ni}).

(b) Repeatabilities for individual studies

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

(c) Meta-analysis: overall repeatability estimates, heterogeneities and publication bias

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

While confidence intervals of mean repeatability estimates (figure 3 and table 1) indicate considerable variability in the repeatability of cognitive performance between studies, inconsistency between effect sizes is better captured by heterogeneity I^2 for meta-analysis [68]. We found moderate to high total heterogeneity ($32\% < I^2 < 88\%$, table 1) as in other across-species meta-analyses [68]. Indeed, a considerable

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

type of R	mean effect size	low CI	high CI	sample size	Egger bias	I^2 total	I^2 laboratory	I^2 species	I^2 experiment
temporal R	0.183	0.088	0.282	40	0.032	0.634	0.063	0.372	0.199
temporal R adjusted for test order	0.150	0.095	0.213	40	0.353	0.661	0.000	0.162	0.499
temporal R adjusted for test order and individual determinants	0.279	0.129	0.428	26	0.399	0.738	0.322	0.000	0.416
contextual R	0.267	0.032	0.477	38	0.194	0.885	0.000	0.725	0.160
contextual R adjusted for test order	0.222	0.129	0.312	32	0.054	0.320	0.000	0.296	0.024
contextual R adjusted for test order and individual determinants	0.195	0.085	0.305	37	0.364	0.593	0.000	0.586	0.007

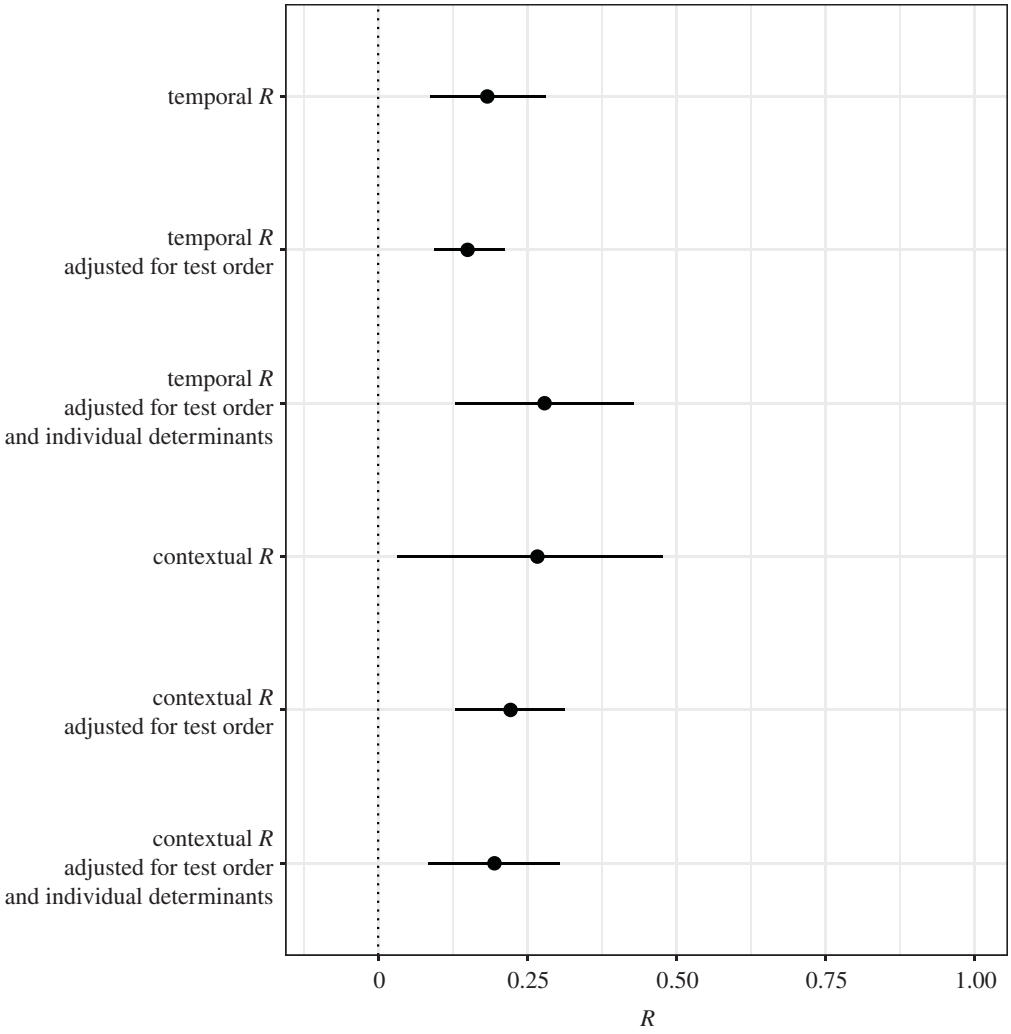

Figure 3. Meta-analytic mean estimates of repeatability (*R*) for temporal and contextual repeatability including unadjusted, adjusted for test order and adjusted for test order plus individual determinants (sex and/or age). We present posterior means and 95% confidence intervals of meta-analyses obtained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

proportion of the total heterogeneity (I^2 total) is due to variations between species (I^2 species). Using repeatability from different cognitive measurements in the same experiment (I^2 experiment) also produced a moderate level of heterogeneity, suggesting that the type of cognitive measurement plays a role in repeatability estimation.

We investigated whether our meta-analysis model showed any bias in publication or selection using a type of Egger's regression. Egger's regressions suggest significant bias for unadjusted temporal *R*. Such bias is probably related to the high number of low sample size studies. To further evaluate the robustness of our mean estimates, we ran a sensitivity analysis using a 'leave one out procedure' (electronic supplementary material, general methods) in which we computed mean estimates by removing a single *R* value for each *R* value in the dataset and generating a distribution of mean estimates. The distribution of 'leave one out' mean estimates was concentrated around the original mean estimate, which suggests that meta-analytic results are not driven by one particular *R* value (electronic supplementary material, figure S10). Finally, we assessed whether mean estimates obtained for each type of *R* analysis were significantly different from each other using multiple *t*-test comparisons. We found that adjusted temporal *R* for test order was significantly lower than other types of *R* analyses before correcting for multiple comparisons (electronic supplementary material, table S5). However, we found no significant differences after correcting for multiple comparisons for all combinations of *R* analyses.

(d) Meta-regression: effects of moderators

To better understand the factors that influence heterogeneity of repeatability, we included the type of cognitive performance measurement, the type of cognitive task, median delay between repetitions, origin of the subjects, experimental context, taxonomic class and publication status as moderators in our models of repeatability. Effects of those factors on raw *R* values can be inspected visually in electronic supplementary material, figures S3–S9. However, to assess the effects of these factors while accounting for variation in sample size and repetition number between studies, meta-analytical tools are necessary. The total number of repeatability values compiled for each type of *R* analysis (table 1) was not sufficient to run a full model to assess the effects of all seven moderators together. We therefore ran seven independent univariate (multilevel) meta-regression models, which revealed that the type of cognitive performance measurement significantly influenced all types of *R* values, except for unadjusted temporal values (table 2), and accounted for 14 to 100% of the variance (R^2_c). The investigation of beta coefficients associated with each type of cognitive measurement (electronic supplementary material, figure S11) suggests that normalized

index (scores computed specifically for the study e.g. Matzel *et al*. dataset) and SUC measures are significantly more repeatable for contextual R_{ni} estimates than other types of R analyses. However, as this pattern is not observed for other types of R analyses, results should be interpreted with caution. Publication status also significantly influenced contextual repeatability and accounted for 24 to 70% of the variance (table 2), with published R values being significantly higher than the R values that are computed from primary data (electronic supplementary material, figure S12).

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

4. Discussion

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

(a) Are measures of cognition repeatable?

High plasticity of cognitive processes may result in low or null estimates of repeatability. Yet, we found a significant, but low, average R estimate for unadjusted temporal repeatability of cognitive performance ($R = 0.18$). Our highest temporal repeatability estimate adjusted for test order and individual determinants reached $R = 0.28$. Although this estimate remains lower than that observed for animal personality and other behaviours (average $R = 0.37$, $R = 0.29$, $R = 0.41$: see [27–29], respectively), our findings suggest that individual variation in performance on the same cognitive task is moderately consistent across time in a wide range of taxa.

Table 2. Summary of meta-regression models. Conditional R^2 (R^2_c) and significance (p-values (p-val.) from omnibus test) of each moderator from the seven univariate meta regressions are presented. n.a., not applicable.

type of R	R^2_c cog. meas.	p-val cog. meas.	R^2_c cog. task	p-val cog. task	R^2_c delay	p-val delay	R^2_c origin	p-val origin	R^2_c exp. context	p-val exp. context	R^2_c class	p-val class	R^2_c pub.	p-val pub.
temporal R	0.407	0.129	0.139	0.879	0.121	0.121	0.188	0.165	0.007	0.739	0.543	0.103	0.000	0.977
temporal R adjusted for test order	0.281	0.013	0.087	0.990	0.023	0.599	0.022	0.617	0.030	0.473	0.352	0.145	0.063	0.202
temporal R adjusted for test order and individual determinants	0.559	0.005	0.285	0.442	0.281	0.031	0.138	0.243	0.015	0.697	0.256	0.246	0.041	0.593
contextual R	0.138	0.002	0.189	0.000	0.004	0.688	0.003	0.823	0.047	0.333	0.153	0.482	0.236	0.039
contextual R adjusted for test order	1.000	0.000	0.919	0.071	0.089	0.496	0.539	0.013	0.153	0.245	0.582	0.445	n.a.	n.a.
contextual R adjusted for test order and individual determinants	0.635	0.000	0.051	0.970	0.047	0.485	0.263	0.092	0.016	0.558	0.020	0.888	0.696	0.001

This result is particularly striking because internal and external influences on task performance are unlikely to be identical between trials; such influences should inflate intra-individual variation between trials, and therefore reduce R. The results we obtained are in line with low to moderate heritability estimates of cognitive performance collected from laboratory populations (reviewed in [69], also see [70,71]) and with selectively bred animals that have shown large differences in, for example, numerical learning in guppies [20], oviposition learning in *Drosophila* [72] and butterflies [73], or maze navigation in rats [74]. These findings may promote future investigation of individual variation in cognitive performance, ideally as a first step towards assessing heritability, the effect of developmental environment and experience on this variation, and examining potential evolutionary consequences of this variation [6,75].

Contextual repeatability was assessed by examining performance on novel variants of the same task (e.g. change of stimuli dimension) or different tasks that we considered assessed the same putative cognitive process. The use of different task variants has been advocated to further improve our understanding of cognitive processes, for instance in the context of assessing convergent validity of tasks ([25,76]). Accordingly, our estimates of contextual repeatability were moderate ($R = 0.20–0.27$) and significant, indicating that the use of different stimuli dimensions, perceptual dimensions, apparatuses and tests allows accurate measures of repeatable variation of individual cognitive performance. However, our interpretation of R values assumes that performance on each cognitive test is independent of other traits that could be repeatable as well, such as motor capacities, motivation or personality traits [25].

Accurate estimates of contextual repeatability may be confounded in tasks that use different stimuli or perceptual dimensions. For instance, adaptive specializations that result in differential attention to particular stimuli may result in high within-individual variation in performance over contexts, or in low between-individual variation in one or both contexts [42] (e.g. individuals of some species may show greater variation in their performance when learning a shape discrimination, but show relatively little variation when learning a colour discrimination, even if both tasks require visual-cue learning e.g. [77,78]). Using different tasks or apparatuses to examine the same putative cognitive process may also lead to low contextual repeatability if the salience of stimuli differs between apparatuses. For example, presenting stimuli on a touchscreen as opposed to presenting stimuli with solid objects may vary the salience of stimuli [79]. Such differences may inflate within-individual variance and thus decrease repeatability. Finally, while we may assume similar cognitive processes are involved in variants of the same task, we may obtain low contextual repeatability if the variants require different cognitive processes. One possible solution is to conduct repeatability analyses on the portion of variance likely due to a shared cognitive process by incorporating measures of 'micro-behaviours'. For example, Chow and colleagues [80] used the response latencies to correct and incorrect stimuli to reflect inhibitory control, and the rate of head-switching (head-turning between stimuli) to reflect attention, alongside using the number of errors in learning a colour discrimination-reversal learning task on a touchscreen. Assessing micro-behaviours may therefore capture specific processes that are more closely related to the general cognitive process than more classical approaches. Accordingly, assays of repeatability of cognitive performances could then be examined by repeatedly recording a suite of micro-behavioural traits as well as traditional measures of performance in the same, or variants of the same, task.

(b) Test order and the repeatability of cognitive performance

Animals may improve their performance with increased learning/experience of the same task or on different but related tasks. Hence, controlling for time-related changes (i.e. the number of repetitions of the same task) or task presentation order (i.e. test order) may produce more accurate estimates of repeatability [81]. However, while our adjusted estimates of temporal and contextual repeatability remained significant when controlling for test order, they did not increase (table 1 and figure 3). These findings suggest that repetition number, or task order, may have a negligible influence on repeatability, at least within the range of values represented in our sample.

Estimates of temporal repeatability (electronic supplementary material, table S1) suggest that there may, however, be an optimal number of repetitions when estimating individual variation in cognitive performance. Indeed, prolonged exposure to the same task may reduce most, if not all, between-individual variation in performance (i.e. individuals reach a plateau in performance with increased experience of the same task): high repetitions of the same task (ranging from 7 to 80 repetitions) produced moderate–low repeatability (mean $R = 0.22$), whereas analyses with low repetitions (ranging from 2 to 3 repetitions) produced a moderate–high repeatability (mean $R = 0.42$). Consequently, increasing the number of measures of cognitive performance strengthens memory and learning on a given task, which may increase within-individual variance between tests as internal and external conditions change across repetitions. Likewise, memory and learning may increase within-individual variance between different tasks owing to carry-over effects. Carry-over effects on repeatability may be controlled by running all tests in the same order for all subjects, and by including test number or test date for a given task [81]. The effect of test order on contextual repeatability should, however, be treated with caution, as it may be influenced by the number of R estimates based on small sample size studies, and may also result from Generalized Linear Mixed Model-based repeatability approaches which force R to be positive, in comparison with unadjusted R. Nevertheless, studying the impact of repetition number or prior test exposure may help improve our understanding of how experience can influence cognitive performance.

(c) Individual determinants of the repeatability of cognitive performance

The addition of individual effects such as sex and age, when available, appeared to increase temporal but not contextual repeatability, relative to models that only included test order (table 1 and figure 3). This effect on temporal repeatability may partly result from differences in the processes that underlie performance on cognitive tasks between juveniles and adults. For example, immature freshwater snails, *Lymnaea stagnalis*, show impaired memory for the association between a light flash and the whole body withdrawal

response until they reach maturity [82], juvenile Australian magpies, *Cracticus tibicen*, show impaired performance on a spatial memory task when tested 100 days after fledging compared with those birds that were tested 200 and 300 days after fledging [15], and honeybee, *Apis mellifera* L., workers show impaired spatial memory when tested under 16 days of age as adults compared with their counterparts that were older than 16 days [83]. Adult Eurasian harvest mice, *Micromys minutus*, also show higher repeatability than juveniles on a spatial recognition task [50]. Controlling for age and developmental life-stage, either experimentally (e.g. targeting one age group) or statistically, may therefore play an important role in obtaining accurate estimates of repeatability of cognitive performance.

Males and females may also experience different selective pressures on given cognitive processes that reflect different fitness consequences. Examples of such sex differences include spatial orientation and reference memory in rodents [84], colour and position cue learning in chicks [85], and foraging innovation in guppies [86]. Sex differences in cognitive processes may result from mating behaviours such as territory defence or mate searching, which may reduce between-individual variation within the same sex. Here, we have only examined and discussed a few of the individual factors that can influence measures of cognitive performance across individuals, and thus potentially impact estimates of repeatability. We suggest that the choice of variables included in analyses of adjusted repeatability should reflect the goals of the study, and include explanations of what aspects are controlled for and, more importantly, why [24].

(d) Moderators of the repeatability of cognitive performance

Variation among studies used in a meta-analysis can cause heterogeneity in effect sizes that is directly attributable to the experimental approach. Accounting for such variation can provide insights into which factors influence the trait of interest [68]. For example, we might expect that repeated measurements that are obtained after shorter time intervals may produce better estimates of repeatability because the internal and external states of individuals may be more similar [27]. However, our results suggest that the interval between two tasks had no influence on most estimates of temporal or contextual repeatability. Although animals may form memory associations on a given test, our findings suggest a negligible influence of carry-over effects on the relative extent of between-versus within-individual variation.

We found that the type of cognitive performance measure had a strong effect on estimates of repeatability (table 2). For contextual repeatability, the lowest estimated R values were obtained for latency measures, with most confidence intervals of estimates overlapping with 0 (electronic supplementary material, figure S11). The low repeatability of latency measures between performance using different apparatuses may result from ceiling effects (e.g. individuals may solve an easy task with similar latencies but show greater variation when solving a more difficult problem) and floor effects (e.g. individuals may use the maximum time that is given in a trial to solve a more difficult problem but show variation for an easy task) [87,88]. Accordingly, the effects of internal or external variables on repeatability may be minimized by using binary measures such as SUC. Our results indicate that certain types of measures (e.g. latency or the number of trials) used in some cognitive tasks are more sensitive to internal or external contextual variables than others and thus provide less reliable measures of R. However, we suggest that moderator effects should be interpreted with caution, as constraints on our sample size prevented us from controlling for other fixed effects when revealing each moderator effect as well as potential interaction effects. Our approach of univariate testing may, therefore, have been more liberal than a full model approach. While our results generally suggest that most moderators did not explain variation in the repeatability of inter-individual variation in cognitive performance across studies, these factors may still be important to consider when designing experiments for a particular species.

(e) General conclusion and future research

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

While we attempt to understand the repeatability of cognitive performance, we acknowledge that this is an emerging and rapidly developing field. Accordingly, this study suffers some limitations, including a modest sample size (both for the number of studies included and for the number of subjects provided in each study), which reduces the robustness of the conclusions regarding the effect of potential moderators. Moreover, this study may also suffer some undetected bias in data collection, as the majority of data were obtained either from colleagues that presented at a workshop on the 'Causes and consequences of individual variation in cognition' or from researchers who work on individual differences known to the workshop participants. However, we argue that the inclusion of unpublished data is a useful approach to gaining a better representation of the true range of repeatabilities, given that we found published studies to provide higher R than unpublished studies. Future studies may, therefore, benefit from the growing body of literature on individual differences in cognition ([42,75,89,90]). Note that other studies collecting repeated measures from repetitions of a same test, or functionally similar tests, could also offer valuable datasets, even when their aim is not the quantification of consistent individual differences. To facilitate future meta-analyses, we suggest that authors of such papers: (i) publish their datasets using the finest-grained information available (e.g. trial-by-trial instead of aggregate values, such as proportion of correct choices or trials); (ii) include information on potential moderators (e.g. date of test, subject's origin) and other fixed effects (e.g. sex, age) that may need to be controlled for; and (iii) include and standardize the term 'cognitive repeatability' in their keywords.

Future avenues for research may include: (1) studying the repeatability of reaction norms of cognitive performance (i.e. its plasticity [44,91] over gradients of interest, for example, deprivation level or housing conditions), so as to assess the generality of the individual differences that are captured by cognitive tasks across different environments and physiological states; and (2) partitioning the variance among and within individuals, by making use of multiple (more than 4) trials recorded for each individual [92]. By partitioning variance in cognitive performance at various hierarchical levels (within and between individuals), we may complement approaches that quantify variation at other levels (populations and species) and hence further our understanding of the evolution of cognition. This approach may provide a greater understanding of the factors that influence repeatability estimates, which are based on a ratio, and thus do not allow the separation of variance that is due to different phenotypes (among-individual) from those due to the plasticity in the response of each animal (within-individual). Separating these values could provide a way to focus on the portion of variance that is expected to be heritable, and to test hypotheses on the factors that affect variation within individuals between repeated trials.

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

Data accessibility. We provide access to the information of general methods (electronic supplementary material) and primary data (https://doi.org/10.6084/m9.figshare.6431549.v1).

Authors' contributions. M.C., P.K.Y.C., J.O.v.H., A.S.C., S.E.G.L. and J.M.-F. defined research; all authors except S.N. contributed primary data either for the initial or for the final manuscript; M.C. conducted analyses and S.N. provided code and commented on analyses; M.C., P.K.Y.C. and J.O.v.H wrote the manuscript with contributions from A.S.C. and J.M.-F. Authors who contributed data wrote their respective methods sections for the supporting information. All authors read and commented on the manuscript.

Competing interests. All authors declare there are no competing interests.

Funding. P.K.Y.C. is supported by the Japan Society for the Promotion of Science (PE1801); J.O.v.H. was funded by an ERC consolidator grant to J. Madden (616474). M.C. as well as this research was supported by a grant from the Human Frontier Science Program to A.S.C. and J.M.-F. (RGP0006/2015) and a Natural Sciences and Engineering Research Council of Canada Discovery Grant to J.M.F. (435596-2013).

References

1. Shettleworth SJ. 2010 *Cognition, evolution, and behavior*. New York, NY: Oxford University Press.
2. van Horik J, Emery NJ. 2011 Evolution of cognition. *Wiley Interdiscip. Rev. Cogn. Sci.* **2**, 621–633. (doi:10.1002/wcs.144)
3. van Horik JO, Clayton NS, Emery NJ. 2012 *Convergent evolution of cognition in corvids, apes and other animals*. New York, NY: Oxford University Press.
4. MacLean EL *et al.* 2012 How does cognition evolve? Phylogenetic comparative psychology. *Anim. Cogn.* **15**, 223–238. (doi:10.1007/s10071-011-0448-8)
5. Thornton A, Isden J, Madden JR. 2014 Toward wild psychometrics: linking individual cognitive differences to fitness. *Behav. Ecol.* **25**, 1299–1301. (doi:10.1093/beheco/aru095)
6. Cauchoix M, Chaine AS. 2016 How can we study the evolution of animal minds? *Front. Psychol.* **7**, 358. (doi:10.3389/fpsyg.2016.00358)
7. Dukas R, Bernays EA. 2000 Learning improves growth rate in grasshoppers. *Proc. Natl Acad. Sci. USA* **97**, 2637–2640. (doi:10.1073/pnas.050461497)
8. Raine NE, Chittka L. 2008 The correlation of learning speed and natural foraging success in bumble-bees. *Proc. R. Soc. B* **275**, 803–808. (doi:10.1098/rspb.2007.1652)
9. Pasquier G, Grüter C. 2016 Individual learning performance and exploratory activity are linked to colony foraging success in a mass-recruiting ant. *Behav. Ecol.* **27**, 1702–1709. (doi:10.1093/beheco/arw079)
10. Maille A, Schradin C. 2016 Survival is linked with reaction time and spatial memory in African striped mice. *Biol. Lett.* **12**, 20160346. (doi:10.1098/rsbl.2016.0346)
11. Kotrschal A, Buechel SD, Zala SM, Corral-Lopez A, Penn DJ, Kolm N. 2015 Brain size affects female but not male survival under predation threat. *Ecol. Lett.* **18**, 646–652. (doi:10.1111/ele.12441)
12. Keagy J, Savard J-F, Borgia G. 2009 Male satin bowerbird problem-solving ability predicts mating success. *Anim. Behav.* **78**, 809–817. (doi:10.1016/j.anbehav.2009.07.011)
13. Cole EF, Morand-Ferron J, Hinks AE, Quinn JL. 2012 Cognitive ability influences reproductive life history variation in the wild. *Curr. Biol.* **22**, 1808–1812. (doi:10.1016/j.cub.2012.07.051)
14. Cauchard L, Boogert NJ, Lefebvre L, Dubois F, Doligez B. 2013 Problem-solving performance is correlated with reproductive success in a wild bird population. *Anim. Behav.* **85**, 19–26. (doi:10.1016/j.anbehav.2012.10.005)
15. Ashton BJ, Ridley AR, Edwards EK, Thornton A. 2018 Cognitive performance is linked to group size and affects fitness in Australian magpies. *Nature* **554**, 364–367. (doi:10.1038/nature25503)
16. Isden J, Panayi C, Dingle C, Madden J. 2013 Performance in cognitive and problem-solving tasks in male spotted bowerbirds does not correlate with mating success. *Anim. Behav.* **86**, 829–838. (doi:10.1016/j.anbehav.2013.07.024)
17. Dunlap AS, Stephens DW. 2016 Reliability, uncertainty, and costs in the evolution of animal learning. *Curr. Opin. Behav. Sci.* **12**, 73–79. (doi:10.1016/j.cobeha.2016.09.010)
18. Mery F. 2013 Natural variation in learning and memory. *Curr. Opin. Neurobiol.* **23**, 52–56. (doi:10.1016/j.conb.2012.09.001)
19. Kawecki TJ. 2009 Evolutionary ecology of learning: insights from fruit flies. *Popul. Ecol.* **52**, 15–25. (doi:10.1007/s10144-009-0174-0)
20. Kotrschal A, Rogell B, Bundsen A, Svensson B, Zajitschek S, Brännström I, Immler S, Maklakov AA, Kolm N. 2013 Artificial selection on relative brain size in the guppy reveals costs and benefits of evolving a larger brain. *Curr. Biol.* **23**, 168–171. (doi:10.1016/j.cub.2012.11.058)
21. Endler JA. 1986 *Natural selection in the wild*. Princeton, NJ: Princeton University Press.
22. Dohm MR. 2002 Repeatability estimates do not always set an upper limit to heritability. *Funct. Ecol.* **16**, 273–280. (doi:10.1046/j.1365-2435.2002.00621.x)
23. Edwards AWF, Falconer DS. 1982 Introduction to quantitative genetics. *Biometrics* **38**, 1128. (doi:10.2307/2529912)
24. Wilson AJ. 2018 How should we interpret estimates of individual repeatability? *Evol. Lett.* **2**, 4–8. (doi:10.1002/evl3.40)
25. Griffin AS, Guillette LM, Healy SD. 2015 Cognition and personality: an analysis of an emerging field. *Trends Ecol. Evol.* **30**, 207–214. (doi:10.1016/j.tree.2015.01.012)
26. Martin JGA, Réale D. 2008 Temperament, risk assessment and habituation to novelty in eastern chipmunks, *Tamias striatus*. *Anim. Behav.* **75**, 309–318. (doi:10.1016/j.anbehav.2007.05.026)
27. Bell AM, Hankison SJ, Laskowski KL. 2009 The repeatability of behaviour: a meta-analysis. *Anim. Behav.* **77**, 771–783. (doi:10.1016/j.anbehav.2008.12.022)
28. Dochtermann NA, Schwab T, Sih A. 2015 The contribution of additive genetic variation to personality variation: heritability of personality. *Proc. R. Soc. B* **282**, 20142201.
29. Holtmann B, Lagisz M, Nakagawa S. 2017 Metabolic rates, and not hormone levels, are a likely mediator

of between-individual differences in behaviour: a meta-analysis. *Funct. Ecol.* **31**, 685–696. (doi:10.1111/1365-2435.12779)

30. Dingemanse N, Réale D. 2005 Natural selection and animal personality. *Behaviour* **142**, 1159–1184. (doi:10.1163/156853905774539445)
31. Nicolaus M, Tinbergen JM, Bouwman KM, Michler SPM, Ubels R, Both C, Kempenaers B, Dingemanse NJ. 2012 Experimental evidence for adaptive personalities in a wild passerine bird. *Proc. R. Soc. B* **279**, 4885–4892. (doi:10.1098/rspb.2012.1936)
32. Dingemanse NJ, Wolf M. 2010 Recent models for adaptive personality differences: a review. *Phil. Trans. R. Soc. B* **365**, 3947–3958. (doi:10.1098/rstb.2010.0221)
33. Dall SRX, Houston AI, McNamara JM. 2004 The behavioural ecology of personality: consistent individual differences from an adaptive perspective. *Ecol. Lett.* **7**, 734–739. (doi:10.1111/j.1461-0248.2004.00618.x)
34. Hughes C, Adlam A, Happé F, Jackson J, Taylor A, Caspi A. 2000 Good test–retest reliability for standard and advanced false-belief tasks across a wide range of abilities. *J. Child Psychol. Psychiatry* **41**, 483–490. (doi:10.1111/1469-7610.00633)
35. Guenther A, Brust V. 2017 Individual consistency in multiple cognitive performance: behavioural versus cognitive syndromes. *Anim. Behav.* **130**, 119–131. (doi:10.1016/j.anbehav.2017.06.011)
36. Brust V, Guenther A. 2017 Stability of the guinea pigs personality–cognition–linkage over time. *Behav. Processes* **134**, 4–11. (doi:10.1016/j.beproc.2016.06.009)
37. Gibelli J, Dubois F. 2016 Does personality affect the ability of individuals to track and respond to changing conditions? *Behav. Ecol.* **28**, 101–107. (doi:10.1093/beheco/arw137)
38. Tello-Ramos MC, Branch CL, Pitera AM, Kozlovsky DY, Bridge ES, Pravosudov VV. 2018 Memory in wild mountain chickadees from different elevations: comparing first-year birds with older survivors. *Anim. Behav.* **137**, 149–160. (doi:10.1016/j.anbehav.2017.12.019)
39. Chittka L, Dyer AG, Bock F, Dornhaus A. 2003 Psychophysics: bees trade off foraging speed for accuracy. *Nature* **424**, 388. (doi:10.1038/424388a)
40. Dalesman S, Rendle A, Dall SRX. 2015 Habitat stability, predation risk and 'memory syndromes'. *Sci. Rep.* **5**, 10538. (doi:10.1038/srep10538)
41. Morand-Ferron J, Cole EF, Quinn JL. 2016 Studying the evolutionary ecology of cognition in the wild: a review of practical and conceptual challenges. *Biol. Rev.* **91**, 367–389. (doi:10.1111/brv.12174)
42. Rowe C, Healy SD. 2014 Measuring variation in cognition. *Behav. Ecol.* **25**, 1287–1292. (doi:10.1093/beheco/aru090)
43. van Horik JO, Langley EJG, Whiteside MA, Laker PR, Beardsworth CE, Madden JR. 2018 Do detour tasks provide accurate assays of inhibitory control? *Proc. R. Soc. B* **285**, 20180150. (doi:10.1098/rspb.2018.0150)
44. Dingemanse NJ, Dochtermann NA. 2013 Quantifying individual variation in behaviour: mixed-effect modelling approaches. *J. Anim. Ecol.* **82**, 39–54. (doi:10.1111/1365-2656.12013)
45. Nakagawa S, Schielzeth H. 2010 Repeatability for Gaussian and non-Gaussian data: a practical guide for biologists. *Biol. Rev.* **85**, 935–956. (doi:10.1111/j.1469-185X.2010.00141.x)
46. Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group. 2009 Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Med.* **6**, e1000097. (doi:10.1371/journal.pmed.1000097)
47. Duckworth AL, Kern ML. 2011 A meta-analysis of the convergent validity of self-control measures. *J. Res. Pers.* **45**, 259–268. (doi:10.1016/j.jrp.2011.02.004)
48. Rodríguez RL, Gloudeman MD. 2011 Estimating the repeatability of memories of captured prey formed by *Frontinella communis* spiders (Araneae: Linyphiidae). *Anim. Cogn.* **14**, 675–682. (doi:10.1007/s10071-011-0402-9)
49. Schuster AC, Zimmermann U, Hauer C, Foerster K. 2017 A behavioural syndrome, but less evidence for a relationship with cognitive traits in a spatial orientation context. *Front. Zool.* **14**, 19. (doi:10.1186/s12983-017-0204-2)
50. Schuster AC, Carl T, Foerster K. 2017 Repeatability and consistency of individual behaviour in juvenile and adult Eurasian harvest mice. *Naturwissenschaften* **104**, 10. (doi:10.1007/s00114-017-1430-3)
51. Shaw RC. 2017 Testing cognition in the wild: factors affecting performance and individual consistency in two measures of avian cognition. *Behav. Processes* **134**, 31–36. (doi:10.1016/j.beproc.2016.06.004)
52. Cole EF, Cram DL, Quinn JL. 2011 Individual variation in spontaneous problem-solving performance among wild great tits. *Anim. Behav.* **81**, 491–498. (doi:10.1016/j.anbehav.2010.11.025)
53. Koricheva J, Gurevitch J, Mengersen K. 2013 *Handbook of meta-analysis in ecology and evolution*. Princeton, NJ: Princeton University Press.
54. R Development Core Team. 2017 *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org
55. Stoffel MA, Nakagawa S, Schielzeth H. 2017 rptR: repeatability estimation and variance decomposition by generalized linear mixed-effects models. *Methods Ecol. Evol.* **8**, 1639–1644. (doi:10.1111/2041-210X.12797)
56. Lessells CM, Boag PT. 1987 Unrepeatable repeatabilities: a common mistake. *Auk* **104**, 116–121. (doi:10.2307/4087240)
57. Holtmann B, Santos ESA, Lara CE, Nakagawa S. 2017 Personality-matching habitat choice, rather than behavioural plasticity, is a likely driver of a phenotype–environment covariance. *Proc. R. Soc. B* **284**, 20170943. (doi:10.1098/rspb.2017.0943)
58. Wolak ME, Fairbairn DJ, Paulsen YR. 2011 Guidelines for estimating repeatability. *Methods Ecol. Evol.* **3**, 129–137. (doi:10.1111/j.2041-210X.2011.00125.x)
59. Viechtbauer W. 2010 Conducting meta-analyses in R with the metafor package. *J. Stat. Softw.* **36**, 1–48. (doi:10.18637/jss.v036.i03)
60. Hinchliff CE *et al.* 2015 Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proc. Natl Acad. Sci. USA* **112**, 12 764–12 769. (doi:10.1073/pnas.1423041112)
61. Michonneau F, Brown JW, Winter D. 2016 rotl, an R package to interact with the Open Tree of Life data. (doi:10.7287/peerj.preprints.1471)
62. Nakagawa S, Schielzeth H. 2012 The mean strikes back: mean–variance relationships and heteroscedasticity. *Trends Ecol. Evol.* **27**, 474–475. (doi:10.1016/j.tree.2012.04.003)
63. Benjamini Y, Hochberg Y. 1995 Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **57**, 289–300.
64. Higgins JPT, Thompson SG. 2002 Quantifying heterogeneity in a meta-analysis. *Stat. Med.* **21**, 1539–1558. (doi:10.1002/sim.1186)
65. Nakagawa S, Santos ESA. 2012 Methodological issues and advances in biological meta-analysis. *Evol. Ecol.* **26**, 1253–1274. (doi:10.1007/s10682-012-9555-5)
66. Egger M, Davey Smith G, Schneider M, Minder C. 1997 Bias in meta-analysis detected by a simple, graphical test. *BMJ* **315**, 629–634. (doi:10.1136/bmj.315.7109.629)
67. Moreno SG, Sutton AJ, Ades AE, Stanley TD, Abrams KR, Peters JL, Cooper NJ. 2009 Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. *BMC Med. Res. Methodol.* **9**, 2. (doi:10.1186/1471-2288-9-2)
68. Nakagawa S, Noble DWA, Senior AM, Lagisz M. 2017 Meta-evaluation of meta-analysis: ten appraisal questions for biologists. *BMC Biol.* **15**, 18. (doi:10.1186/s12915-017-0357-7)
69. Croston R, Branch CL, Kozlovsky DY, Dukas R, Pravosudov VV. 2015 Heritability and the evolution of cognitive traits. *Behav. Ecol.* **26**, 1447–1459. (doi:10.1093/beheco/arv088)
70. Sauce B, Bendrath S, Herzfeld M, Siegel D, Style C, Rab S, Korabelnikov J, Matzel LD. 2018 The impact of environmental interventions among mouse siblings on the heritability and malleability of general cognitive ability. *Phil. Trans. R. Soc. B* **373**, 20170289. (doi:10.1098/rstb.2017.0289)
71. Sorato E, Zidar J, Garnham L, Wilson A, Lovlie H. 2018 Heritabilities and co-variation among cognitive traits in red junglefowl. *Phil. Trans. R. Soc. B* **373**, 20170285. (doi:10.1098/rstb.2017.0285)
72. Burger JMS, Kolss M, Pont J, Kawecki TJ. 2008 Learning ability and longevity: a symmetrical evolutionary trade-off in *Drosophila*. *Evolution* **62**, 1294–1304. (doi:10.1111/j.1558-5646.2008.00376.x)
73. Snell-Rood EC, Davidowitz G, Papaj DR. 2011 Reproductive tradeoffs of learning in a butterfly. *Behav. Ecol.* **22**, 291–302. (doi:10.1093/beheco/arq169)
74. Tryon RC. 1940 Studies in individual differences in maze ability. VII. The specific components of maze

ability, and a general theory of psychological components. *J. Comp. Psychol.* **30**, 283–335. (doi:10.1037/h0054238)

75. Thornton A, Lukas D. 2012 Individual variation in cognitive performance: developmental and evolutionary perspectives. *Phil. Trans. R. Soc. B* **367**, 2773–2783. (doi:10.1098/rstb.2012.0214)
76. Völter CJ, Tinklenberg B, Call J, Seed AM. 2018 Comparative psychometrics: establishing what differs is central to understanding what evolves. *Phil. Trans. R. Soc. B* **373**, 20170283. (doi:10.1098/rstb.2017.0283)
77. Wäckers FL, Lewis WJ. 1999 A comparison of color-, shape- and pattern-learning by the hymenopteran parasitoid *Microplitis croceipes*. *J. Comp. Physiol. A* **184**, 387–393. (doi:10.1007/s003590050337)
78. Aronsson M, Gamberale-Stille G. 2008 Domestic chicks primarily attend to colour, not pattern, when learning an aposematic coloration. *Anim. Behav.* **75**, 417–423. (doi:10.1016/j.anbehav.2007.05.006)
79. O'Hara M, Huber L, Gajdon GK. 2015 The advantage of objects over images in discrimination and reversal learning by kea, *Nestor notabilis*. *Anim. Behav.* **101**, 51–60. (doi:10.1016/j.anbehav.2014.12.022)
80. Chow PKY, Leaver LA, Wang M, Lea SEG. 2017 Touch screen assays of behavioural flexibility and error characteristics in Eastern grey squirrels (*Sciurus carolinensis*). *Anim. Cogn.* **20**, 459–471. (doi:10.1007/s10071-017-1072-z)
81. Biro PA, Stamps JA. 2015 Using repeatability to study physiological and behavioural traits: ignore time-related change at your peril. *Anim. Behav.* **105**, 223–230. (doi:10.1016/j.anbehav.2015.04.008)
82. Ono M, Kawai R, Horikoshi T, Yasuoka T, Sakakibara M. 2002 Associative learning acquisition and retention depends on developmental stage in *Lymnaea stagnalis*. *Neurobiol. Learn. Mem.* **78**, 53–64. (doi:10.1006/nlme.2001.4066)
83. Ushitani T, Perry CJ, Cheng K, Barron AB. 2016 Accelerated behavioural development changes fine-scale search behaviour and spatial memory in honey bees (*Apis mellifera* L.). *J. Exp. Biol.* **219**, 412–418. (doi:10.1242/jeb.126920)
84. Jonasson Z. 2005 Meta-analysis of sex differences in rodent models of learning and memory: a review of behavioral and biological data. *Neurosci. Biobehav. Rev.* **28**, 811–825. (doi:10.1016/j.neubiorev.2004.10.006)
85. Vallortigara G. 1996 Learning of colour and position cues in domestic chicks: males are better at position, females at colour. *Behav. Processes* **36**, 289–296. (doi:10.1016/0376-6357(95)00063-1)
86. Laland KN, Reader SM. 1999 Foraging innovation in the guppy. *Anim. Behav.* **57**, 331–340. (doi:10.1006/anbe.1998.0967)
87. Griffin AS, Guez D. 2014 Innovation and problem solving: a review of common mechanisms. *Behav. Processes* **109**, 121–134. (doi:10.1016/j.beproc.2014.08.027)
88. van Horik JO, Madden JR. 2016 A problem with problem solving: motivational traits, but not cognition, predict success on novel operant foraging tasks. *Anim. Behav.* **114**, 189–198. (doi:10.1016/j.anbehav.2016.02.006)
89. Morand-Ferron J, Quinn JL. 2015 The evolution of cognition in natural populations. *Trends Cogn. Sci.* **19**, 235–237. (doi:10.1016/j.tics.2015.03.005)
90. Dougherty LR, Guillette LM. 2018 Linking personality and cognition: a meta-analysis. *Phil. Trans. R. Soc. B* **373**, 20170282. (doi:10.1098/rstb.2017.0282)
91. Martin JGA, Nussey DH, Wilson AJ, Réale D. 2011 Measuring individual differences in reaction norms in field and experimental studies: a power analysis of random regression models. *Methods Ecol. Evol.* **2**, 362–374. (doi:10.1111/j.2041-210X.2010.00084.x)
92. van de Pol M, Wright J. 2009 A simple method for distinguishing within- versus between-subject effects using mixed models. *Anim. Behav.* **77**, 753–758. (doi:10.1016/j.anbehav.2008.11.006)