ESM details and information of individual studies for *'The repeatability of cognitive performance: a meta-analysis'* by Cauchoix M, Chow PKY, van Horik JO, Atance CM**,** Barbeau EJ, Barragan-Jason G, Bize P, Boussard A, Buechel SD, Cabirol A, Cauchard L, Claidière N**,** Dalesman S, Devaud JM, Didic M, Doligez B, Fagot J, Fichtel C, Henke-von der Malsburg J, Hermer E, Huber L, Huebner F, Kappeler PM, Klein S, Langbein J, Langley EJG, Lea SEG, Lihoreau M, Lovlie H, Matzel LD, Nakagawa S, Nawroth C, Oesterwind S, Sauce B, Smith E, Sorato E, Tebbich S, Wallis LJ, Whiteside MA, Wilkinson A, Chaine AS, Morand-Ferron J

List of datasets:

1. **SI Atance, Metcalf, & Thiessen**
2. **SI Barbeau et Didic**
3. **SI Barragan-Jason Gladys**
4. **SI Cabirol et al. 2017**
5. **SI Cauchard, Bize, Grimmond, & Doligez**
6. **SI Cauchard, Boogert, Angers & Doligez**
7. **SI Cauchoix et al.**
8. **SI Chow et al. 2015**
9. **SI Chow et al. 2016**
10. **SI Chow et al. 2017a**
11. **SI Chow et al. 2017b**
12. **SI Chow (unpublished data)**
13. **SI Chow et al. 2018**
14. **SI Claidière et al.**
15. **SI Dalesman 2015**
16. **SI Henke-von der Malsburg & Fichtel 2015**
17. **SI Henke-von der Malsburg & Fichtel 2016**
18. **SI Huebner & Kappeler 2018**
19. **SI Klein et al. 2017**
20. **SI Langley & Whiteside (unpublished data)**
21. **SI Lihoreau et al. 2012a**
22. **SI Lihoreau et al. 2012b**
23. **SI Lihoreau et al. 2011**
24. **SI Matzel et al. Attention dataset**
25. **SI Matzel et al. Learning dataset**
26. **SI Nawroth et al. 2013**
27. **SI Nawroth et al. 2013-14a**
28. **SI Nawroth et al. 2015**
29. **SI Nawroth et al. 2013-14b**
30. **SI Oesterwind et al. (unpublished data)**
31. **SI Sorato & Lovlie**
32. **SI van Horik & Emery**
33. **SI van Horik & Madden 2016**
34. **SI Wallis et al. 2016**
35. **SI Wilkison (unpublished)**

1. SI Atance, Metcalf, & Thiessen

Authors' names and affiliation
Atance C, Metcalf JL, & Thiessen AJ

Authors' affiliation
School of Psychology, University of Ottawa, Canada

Methods for working memory. We administered Backward Digit Span Method to 85 3- to 5-year-olds and was adapted from Davis and Pratt (1995) and more recently by Carlson et al. (2002). Children were asked to repeat a list of single-digit numbers in reverse order. The experimenter used a puppet to demonstrate saying digits backwards: "This is my friend Johnny. Whenever I say numbers, Johnny says them backwards. Listen: '5-8'". In response, Johnny said: "8-5". Children were then asked to do as Johnny had done. They were given a pair of two-digit practice trials and corrected up to two times per practice trial if necessary. The experimenter then proceeded to administer the pairs of test trials. The number of digits on the list increased with each successful performance on both trials in a pair (2, 3, 4, 5, and 6 digits). Children received one point per correct trial and the task was discontinued when they had failed both trials in the same pair. Possible scores on this task ranged from 0 to 10.

References

Carlson, S. M., Moses, L. J., & Breton, C. (2002). How specific is the relation between executive functioning and theory of mind? Contribution of inhibitory control and working memory. Infant and Child Development, 11, 73–92. http://dx.doi.org/10.1002/icd.298

Davis, H. L., & Pratt, C. (1995). The development of children's theory of mind: The working memory explanation. Australian Journal of Psychology, 47, 25–31. http://dx.doi.org/10.1080/00049539508258765

Counting and Labelling Method. This task was administered to 83 3- to 5-year-olds and was designed to measure dual-task performance (Carlson et al., 2002; Gordon & Olson, 1998). The Experimenter (E) presented children with three toys (a horse, a plastic donut, and Play-doh), and named and pointed to each toy in turn. Next, the E counted while pointing to each toy, "1, 2, 3." Finally, the E enumerated and stated the name for each of the three toys, "one is a horse, two is a donut, and three is Play-doh." Children were then given three toys of their own (a turtle, a plastic banana, and a dinosaur) and instructed to repeat the steps the E had performed (i.e., enumerate, label, and then enumerate and label the toys). Children then received a second trial of the task with a new set of items (a plastic orange, a toy car, and a crayon). Children received one point for each correct response to step 3. Possible scores on this task ranged from 0 to 2.

References

Carlson, S. M., Moses, L. J., & Breton, C. (2002). How specific is the relation between

executive functioning and theory of mind? Contribution of inhibitory control and working memory. Infant and Child Development, 11, 73–92. http://dx.doi.org/10.1002/icd.298

Gordon, A. C., & Olson, D. R. (1998). The relation between acquisition of a theory of mind and the capacity to hold in mind. Journal of Experimental Child Psychology, 68, 70–83. http://dx.doi.org/10.1006/jecp.1997.2423

Inhibition was measured using the Grass Snow Method (Carlson & Moses, 2001)
Eighty-six 3- to 5-year-olds were asked to place their hands on top of two felt hand shapes situated beneath a white card and a green card on the table. The experimenter (E) asked children to state the color of grass (green) and the color of snow (white). The E then explained that they were going to play a silly game in which children had to point to the white card when the E said "grass" and to the green card when the E said "snow." There were two practice trials and 16 test trials administered consecutively. Children's first responses to each test trial were scored, even if they self-corrected. Children received a score of 1 on a test trial if they said "grass" in response to a white card or "snow" in response to a green card, and a score of 0 if they said "grass" in response to a green card or "snow" in response to a white card. Total possible scores on this task ranged from 0 to 16.

References

Carlson, S. M., & Moses, L. J. (2001). Individual differences in inhibitory control and children's theory of mind. Child Development, 72, 1032–1053. http://dx.doi.org/10.1111/1467-8624.00333

Davis, H. L., & Pratt, C. (1995). The development of children's theory of mind: The working memory explanation. Australian Journal of Psychology, 47, 25–31. http://dx.doi.org/10.1080/00049539508258765

Whisper Method (Carlson, 2005). This task included 86 3- to 5-year-olds. The experimenter (E) first asked children if they could whisper their names. Children were then told that they would be shown some pictures of cartoon characters and asked to whisper their names. The E then presented a series of 10 cards depicting the cartoon characters, six of which were intended to be familiar to the child (i.e., Shrek, SpongeBob, Dora The Explorer, Buzz Lightyear, Spiderman, and Diego) and four unfamiliar (i.e., Darkwing Duck, Elmer Fudd, Bullwinkle, and Tasmanian Devil). On each familiar character trial children received a score of 2 if they whispered the name, 1 if they spoke in a normal or mixed voice (i.e., if they started in one mode of voice and changed to another as in shouting to whispering or whispering to shouting), and a score of 0 if they shouted out the name. Unfamiliar characters were included in the series so that children would become more excited (and therefore more likely to shout out the name) when presented with a familiar character. Total possible scores on this task ranged from 0 to 12 (2 × 6 familiar characters). Because some children did not recognize all of the "familiar" characters, mean scores were used in the final analyses (i.e., children who only recognized 5 of the "familiar" characters but whispered all of their names received a mean score of 10/5 = 2). Possible mean scores ranged from 0 to 2.

Carlson, S. M. (2005). Developmentally sensitive measures of executive function in preschool children. Developmental Neuropsychology, 28, 595–616. http://dx.doi.org/10.1207/s15326942dn2802_3

Methods for Hanson, Atance, & Paluck (2014)

Working memory was assessed using the Backward Digit Span Method.
This task was administered to 87 3- to 5-year-olds and followed the methods of Carlson (2005), derived from Davis and Pratt (1995). Children were first introduced to a doll (‘‘Jenny’’) and told that Jenny says everything the experimenter (E) says but says it backwards. The E then demonstrated by saying ‘‘5–8’’ and making Jenny say ‘‘8–5.’’ Children were given two practice trials with feedback, followed by two test trials, each with an increasing number of digits beginning with two digits. The task ended when children erred on both trials of a given level. Children were awarded a score of 1 for each successful trial (total score: range = 0–5).

References

Carlson, S. M. (2005). Developmentally sensitive measures of executive function in preschool children. Developmental Neuropsychology, 28, 595–616. http://dx.doi.org/10.1207/s15326942dn2802_3

Davis, H. L., & Pratt, C. (1995). The development of children’s theory of mind: The working memory explanation. Australian Journal of Psychology, 47, 25–31. http://dx.doi.org/10.1080/00049539508258765

Counting and Labelling Method. In this task used by Carlson (2005), the experimenter (E) showed children (N = 90 3- to 5-year-olds) three small two-dimensional wooden objects (e.g., a boat, an apple, and a bird) and children watched while the E labeled them (‘‘boat, apple, bird’’). The experimenter then counted the objects out loud (‘‘one, two, three’’). Finally, the E counted and labeled them each in turn (e.g., ‘‘one is a boat, two is an apple, and three is a bird’’). Children were then asked to complete all three steps in two test trials using different objects. Children’s ability to correctly count and label was awarded a score of 1 for each trial (total score: range = 0–2).

Carlson, S. M. (2005). Developmentally sensitive measures of executive function in preschool children. Developmental Neuropsychology, 28, 595–616. http://dx.doi.org/10.1207/s15326942dn2802_3

2. SI Barbeau et Didic

Authors' names and affiliation:
Emmanuel J. BARBEAU, Centre de recherché Cerveau et Cognition, UPS-UMR5549, Toulouse France and Mira DIDIC, AP-HP Timone and Intitute of Systems neuroscience, Marseille, France

Author's contribution:
EB and MD planned the study and collected and analysed the data. EB and MD wrote the SI methods.

Phil trans authors requested:
Emmanuel Barbeau and Mira Didic

Data sharing (full dataset or summary data):
Full datasets will be shared.

Methods. 42 healthy aged subjects were recruited to participate as control subjects in a study on the predementia stage of Alzheimer's disease (Barbeau et al, 2008, 2012; Didic et al, 2011, 2013). They underwent a set of neuropsychological tests aiming at assessing different cognitive domains, with a special focus on memory. 33 accepted to come back for a follow up study 2.36 years later (SD: 0.64) during which they underwent the same set of tests. These 33 subjects (age at inclusion: 66.76, SD: 7.36; 16 women; number of years of education: 12.16, SD: 3.31) were thus included in the present study and their performance on the same neuropsychological tests were analysed.
These test included measures of:

- Working memory [name of the test: digit span subtest from the WAIS-III. digit span forward, digit span backward. Performance is expressed as the maximum digit span achieved].
- Lexical fluency [name of the test: tâche de Cardebat. words starting with the letter "P" in 2 minutes, repetitions during this task. Performance expressed in number of words correct or repeated].
- Recognition memory for faces [name of the test: Face recognition subtest from the WAIS-III. immediate recognition, delayed recognition. Performance expressed as scaled scores].
- Recognition memory for words [name of the test: Mots25. Performance expressed as percentage of correct recognition].
- Verbal declarative memory test [name of the test: Free and Cued Selective reminding test. Total number of words immediately freely recalled on three consecutive trials (max. 48). Total words recalled when cued. Total words freely recalled after a delay of 20 min. (max. 16). Total intrusions produced during the first three recall trials. Performances expressed as numbers of words].
- Semantic memory for famous events [name of the test: EVE10. Performance on the free recall part of the test. Total performance on the test (max. 60).

- Semantic memory for famous faces [name of the test: VisCel. Total number of faces correctly named (max. 40)].

References

Barbeau EJ, Ranjeva JP, Didic M, Confort-Gouny S, Felician O, Soulier E, Cozzone PJ, Ceccaldi M, Poncet M. (2008). Profile of memory impairment and gray matter loss in amnestic mild cognitive impairment. Neuropsychologia. 7;46(4):1009-19.

Barbeau EJ, Didic M, Joubert S, Guedj E, Koric L, Felician O, Ranjeva JP, Cozzone P, Ceccaldi M. (2012). Extent and neural basis of semantic memory impairment in mild cognitive impairment. J Alzheimers Dis. 28(4):823-37.

Didic M, Barbeau EJ, Felician O, Tramoni E, Guedj E, Poncet M, Ceccaldi M. (2011). Which memory system is impaired first in Alzheimer's disease? J Alzheimers Dis. 27(1):11-22.

Didic M, Felician O, Barbeau EJ, Mancini J, Latger-Florence C, Tramoni E, Ceccaldi M. (2013). Impaired visual recognition memory predicts Alzheimer's disease in amnestic mild cognitive impairment. Dement Geriatr Cogn Disord. 35(5-6):291-9.

3. SI Barragan-Jason Gladys

Author's name and affiliation:
Barragan-Jason Gladys

Institute for Advanced Study in Toulouse, Toulouse School of Economics, 21 allée de Brienne, 31015 Toulouse, France

Author's contribution:
GBJ planned the study, collected the data and wrote the SI methods.

Phil trans authors requested:
Gladys Barragan-Jason

Data sharing (full dataset or summary data):
Full datasets will be shared.

General methods. Forty-three 3-to-8 year-old children engaged in one delay choice task and one delay maintenance task in two distinct testing sessions. In both testing sessions, a female experimenter tested the child individually in a dedicated room in schools and kindergarten located in the district of Saint-Girons in Ariege (South of France).

During the first session, the children performed a delay choice task (e.g., Prencipe & Zelazo, 2005). In this task, children were asked to choose between a small reward now (1 sticker) and a larger reward available after a one-minute delay (2 stickers). If the child selects the one sticker option, he/she can stick it on a background immediately, if he/she chooses the later option, he/she has to wait until the hourglass is finished (after 1 minute) before using the two stickers. Five trials were performed.

During the second session (about a week later), children performed a delay maintenance task (e.g., Mischel, Shoda, & Rodriguez, 1989). In this task, the children were asked to choose between one attractive toy immediately available or two attractive toys available after 10 minutes. A bell was provided allowing them to ring it if they want to change their mind. Delay choice tasks allow us to evaluate the number of trials for which children choose to delay and maintenance tasks allow researchers to evaluate the length of time children will wait for a larger reward.

References

Prencipe, A., & Zelazo, P. D. (2005). Development of affective decision making for self and other: evidence for the integration of first-and third-person perspectives. Psychological Science, 16(7), 501-505. DOI:10.1111/j.0956-7976.2005.01564.x

Mischel, W., Shoda, Y., & Rodriguez, M. I. (1989). Delay of gratification in children. Science, 244(4907), 933–938. DOI: 10.1126/science.2658056

4. SI Cabirol et al. 2017

Author's name and affiliation:
Amélie Cabirol[1,2], Rufus Brooks[1], Claudia Groh[3], Andrew B Barron[2], Jean-Marc Devaud[1]

[1]Research Center on Animal Cognition (CRCA), Center of Integrative Biology (CBI), University of Toulouse; CNRS, UPS, France
[2]Department of Biological Sciences, Macquarie University, North Ryde, NSW 2109, Australia
[3]Department of Behavioral Physiology and Sociobiology, University of Würzburg, Biozentrum, Am Hubland, 97074 Würzburg, Germany

Author's contribution:
AC and JMD conceived the study and designed the methodology; AC collected the data; AC and RB analysed the data; AC, CG, ABB and JMD wrote the manuscript.

Phil trans authors requested:
Amélie Cabirol and Jean-Marc Devaud

Data sharing (full dataset or summary data):
Full datasets (Cabirol_bee_reversal ; Cabirol_bee_differential) will be shared.

General methods. We collected data on olfactory reversal learning (Experiment 1) and two consecutive olfactory differential learning (Experiment 2) in honey bees (Apis mellifera) using the conditioning of the proboscis extension response (PER).

Experimental setup. After emergence from pupae, bees were marked with a dot of colour on the thorax in order to identify their birth date. After nine days, marked bees were collected inside the hive, immobilized on ice and harnessed in small metal tubes allowing movements of their antennae and mouthparts only. Bees were fed with sucrose solution (50% w/w) and left in darkness at room temperature. The cognitive test was performed on the following day.

Cognitive test. Experiment 1 (dataset: Cabirol_bee_reversal). The cognitive performance of 40 honey bees was assessed in an olfactory reversal learning task (figure 1A). This task is based on the conditioning of the proboscis extension response, which is automatically triggered when the antennae of a bee are touched with sucrose solution, thereby allowing the bee to drink the sucrose solution. By presenting an odour shortly before the presentation of sucrose solution, bees can learn the association between the odour and the food reward (positive reinforcement). Successful learning is reflected by PER to the odour alone.

In the reversal learning task, bees had to solve a temporal ambiguity between two learning phases. In the first phase, an odour A was reinforced with sucrose solution while an odour B was unreinforced (A+ vs. B-). In the second phase, one hour later, the odour B became reinforced with sucrose, while the odour A was not reinforced anymore (A- vs. B+). Each phase was composed of 5 reinforced and 5 unreinforced trials in a semi-random order. For each

trial, the odour was presented for 4s and the sucrose solution for 3s (reinforced trials), with a 1s-overlap between the two presentations (trial duration of 40s, inter-trial interval of 8min). Limonene and Eugenol were used alternately as odours A and B (all pure, from Sigma-Aldrich) for different bees.

Experiment 2 (dataset: Cabirol_bee_differential). The cognitive performance of 47 honey bees was assessed in two consecutive olfactory differential learning tasks (figure 1B). The first task was similar to the first phase of reversal learning described above (A+ vs. B-). The second task, one hour later, was similar to the first one but two new odours were used as the reinforced and unreinforced stimuli (C+ vs. D-). 1-nonanol and 1-heptanal were used alternately as odours C and D (all pure, from Sigma-Aldrich) for different bees.

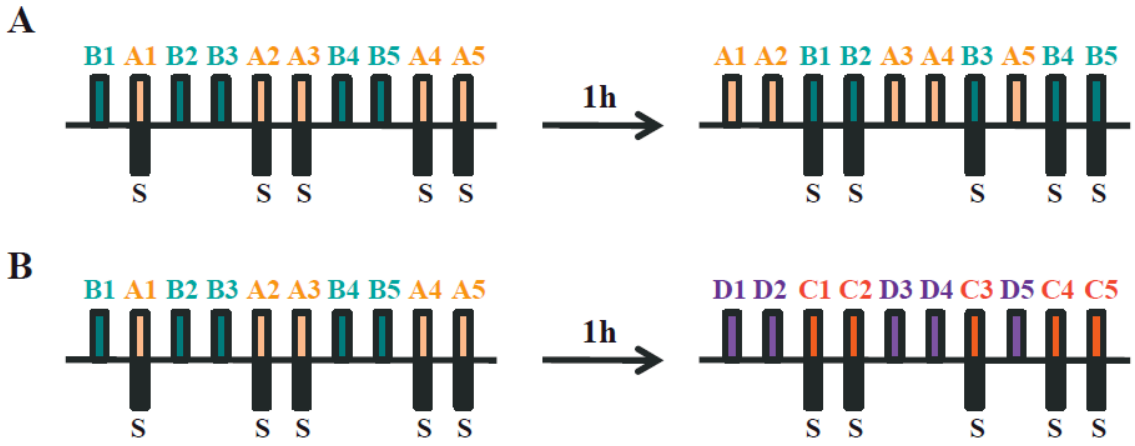


Figure 1. Sequence of presentation of the reinforced and unreinforced odours in the reversal learning task (A) and in two consecutive differential learning tasks (B). Each phase of reversal learning and of the differential learning task is composed of 5 presentations of two odours (A and B, or C and D) in a semi-random order (5 learning trials). One of the two odours presented is reinforced with sucrose (S). In reversal learning (A), the odour A is reinforced in the first phase, not the odour B, and it is the opposite in the second phase one hour later. The first differential learning task (B) is similar to the first phase of reversal learning. In the second differential learning task, one hour later, the odour C is reinforced in the first phase, not the odour D.

Cognitive performance:
The presence or absence of PER in response to the odours was noted 1 or 0 respectively. The ability of each individual bee to solve the tasks within the 5 learning trials was assessed using a learning score calculated as the difference between its response to the reinforced odour and its response to the unreinforced odour in the last trial.

Reference
Cabirol A, Brooks R, Groh C, Barron AB, Devaud JM (2017). Experience during early adulthood shapes the learning capacities and the number of synaptic boutons in the mushroom bodies of honey bees (Apis mellifera). Learning & Memory 24:557-562

5. SI Cauchard, Bize, Grimmond, & Doligez

Author's name and affiliation:
Laure Cauchard[1], Pierre Bize[2], Katie Grimmond[2], Grégory Daniel[3,4], Blandine Doligez[3]

[1]Département de Sciences Biologiques, Université de Montréal, Pavillon Marie-Victorin, bureau D-221, C.P. 6128, succ. Centre-ville, Montréal, Québec, H3C 3J7, Canada
[2]Institute of Biological and Environmental Sciences, University of Aberdeen, Aberdeen, U.K.
[3]CNRS UMR 5558, Université Lyon 1, Université de Lyon, Department of Biometry and Evolutionary Biology, Villeurbanne, France
[4]Animal Ecology, Department of Ecology and Genetics, Evolutionary Biology Centre, Uppsala University, Uppsala, Sweden

Author's contribution:
LC, GD and BD planned the study and collected the data in the field, LC, PB and KG analysed the data, LC, PB and BD wrote the SI methods.

Phil trans authors requested:
Pierre Bize, Laure Cauchard, Blandine Doligez

Data sharing (full dataset or summary data):
Full dataset will be shared

Methods. We collected data on problem-solving and learning in a free-living population of collared flycatchers (Ficedula alibicolis) on the island of Gotland, Sweden, in springs 2012 to 2015. Problem-solving was assessed using a task which was motivated by access to offspring during the peak period of nestling provisioning. The task used here was similar to the one developed for great tits (Parus major) by Cauchard et al. (2013), with the opening system adjusted to the physical abilities of the species (contrary to great tits, collared flycatchers cannot pull a string with a leg).

The task apparatus was attached to the nest-box and consisted of a trap door and 3 levers side-by-side (Figure 1). The door remained closed, blocking the entrance to the nest-box, until the correct lever has been landed on. Either the left or the right lever opened the door when landed on; the middle lever never opened the door. When a parent landed on the correct lever, it could then enter the nest-box and the door closed behind it. The bird could exit from inside the nest-box by pushing the door open. The correct lever remained constant during a trial. We ran two consecutive trials when nestlings were 5 and 6 days of age (trials 1 and 2, respectively), which coincides with the beginning of the peak of nestlings' food demand and parental provisioning rate, to measure learning (trial 1) and reversal learning (trial 2) abilities. To do so, we swapped side of the correct (opening) lever between trial 1 and trial 2: if the left lever opened the door in the learning task (trial 1), the right lever opened the door in the reversal learning task (trial 2).

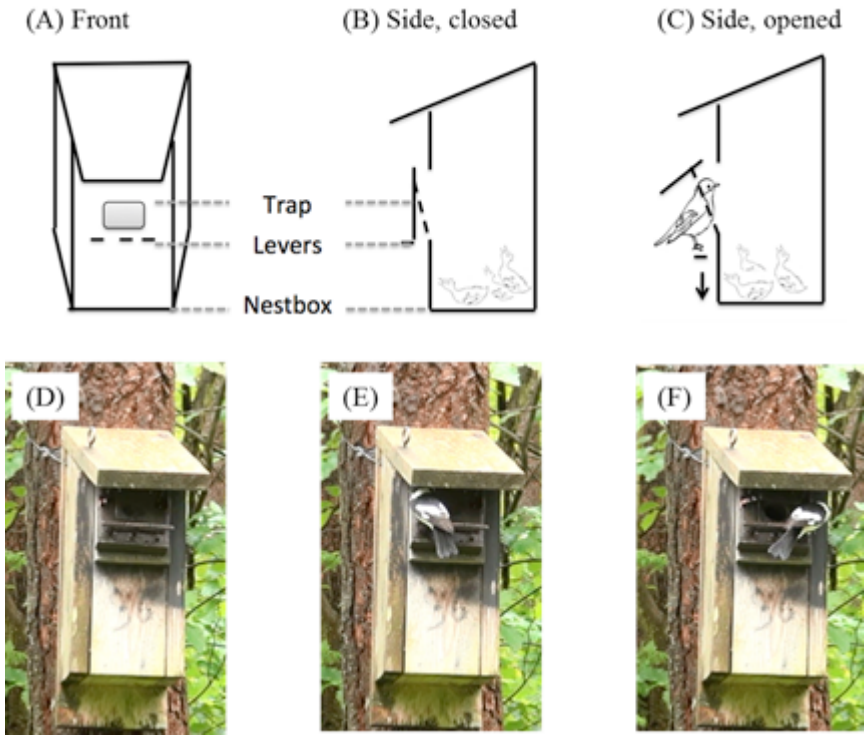


Figure 1. Illustration of the problem-solving task used in free-living adult collared flycatchers. The task apparatus was made of a door blocking the entrance to the nest-box that can be opened by pushing down a lever. Three levers side by side were presented to the birds but only one (either on the right or on the left) was opening the door. Panels A and B provides front and side views of the nest-box and apparatus, door closed. Panel C is a side view, door opened. Panel D shows the task apparatus attached to the nest-box, door closed. Panel E shows an adult male landing on the incorrect lever (here, left) and door remaining closed. Panel F shows a male collared flycatcher landing on the correct lever (here, right) and door opening. We recorded the latency to solve the task and enter the nest-box. The door closed behind the birds when entering the task.

To let the birds habituate to the task apparatus, a dummy apparatus with a permanently open door was attached to the nest-box on the day preceding the first trial. For each trial, the task apparatus was attached in front of the nest-box entrance (replacing the dummy apparatus) for approx. 1 hour between 07:00 AM and 06:30 PM. To prevent starvation if parents were unable to solve the task (i.e. to enter the nest-box), the task was not attached if the nestlings were begging strongly, indicating high hunger level, and the length of each trial was limited to approx. 1 hour Just before attaching the task to the nest-box, we installed a camouflaged video recorder at a distance of approximately 6m in front of the nest-box and recorded the parents' behaviour. All the movements and interactions of each parent with the nest-box and the task were subsequently scored from video recordings (males and females can be discriminated easily based on plumage colour dimorphism).
As in Cauchard et al. (2013), we recorded the latency to solve the task (i.e. enter in the nest-box) as the time elapsed between the first contact with the right lever (i.e. that caused a movement of the door) and the bird's entry into the nest-box, for each parent. When individuals

left the box after contacting the task and then returned with the same attempt, we excluded the time spent away (i.e. the latency only accounted for the time spent trying to enter the nest-box). The results of trial 2 were only analysed for individuals that solved trial 1.

Reference

Cauchard, L., Boogert, N.J., Lefebvre, L., Dubois, F. & Doligez, B. 2013. Problem-solving performance is correlated with reproductive success in a wild bird population. Animal Behaviour 85: 19-26.

6. SI Cauchard, Boogert, Angers & Doligez

Authors' names and affiliation:
Laure Cauchard[1], Neeltje J Boogert[2], Bernard Angers[1], Blandine Doligez[3]

[1]Département de Sciences Biologiques, Université de Montréal, Montreal, QC, Canada.
[2]School of Psychology, University of St Andrews, St Andrews, U.K.
[3]CNRS, Université of Lyon 1, Department of Biometry and Evolutionary Biology, Villeurbanne, France.

Author's contribution:
LC, NJB, BA and BD planned the study. LC collected and analysed the data. LC and BD wrote the SI methods.

Phil trans authors requested:
Laure Cauchard, Blandine Doligez

Data sharing (full dataset or summary data):
Full datasets will be shared

Methods. Data were collected in a population of great tits breeding on the island of Gotland, Sweden (57°10'N, 18°20'E), between April and June 2010 to 2014. This population has been monitored for problem-solving performance during previous studies (Cauchard et al., 2017; Cauchard, Boogert, Lefebvre, Dubois, & Doligez, 2013).

Problem-solving performance was measured directly in the field during breeding. The task consisted of a door placed in front of the entrance of the nest box (see Cauchard et al., 2013 for a complete description of the task). The door was by default closed. To enter, parents had to pull a string placed below the door using their feet to open it and then slip their body under the door. The door then closed automatically behind the bird, but could be simply pushed open from inside by parents to get out. The test was conducted during the peak of nestling food demand (i.e. when nestlings were 7 to 9 days old, between 07:00 AM and 04:00 PM), only when nestlings were satiated (e.g. not begging intensely at the beginning of the test). To avoid nestling starvation if parents were not able to solve the task, the test lasted 1h but was repeated on two consecutive days. We randomly selected breeding pairs to be tested among pairs separated by at least 200 m from the nearest neighbours previously tested, to avoid social learning.

All the movements and interactions of parents with the task were recorded using a camouflaged video recorder at a distance of approx. 6m in front of the nest box. Individuals who succeeded in solving the task (i.e. opening the door and entering the box) were considered to be solvers, while those who contacted the nest box but failed to enter were considered to be non-solvers (i.e. we defined problem-solving status as a binary variable). Behavioural experiments were authorized by the Swedish Committee for Experiments on Animals and conducted in

accordance with international standards on animal welfare as well as being compliant with local and national regulations.

References

Cauchard, L., Angers, B., Boogert, N. J., Lenarth, M., Bize, P., & Doligez, B. (2017). An Experimental Test of a Causal Link between Problem-Solving Performance and Reproductive Success in Wild Great Tits. [Original Research]. Frontiers in Ecology and Evolution, 5(107).

Cauchard, L., Boogert, N. J., Lefebvre, L., Dubois, F., & Doligez, B. (2013). Problem-solving performance is correlated with reproductive success in a wild bird population. Animal Behaviour, 85(1), 19-26.

7. SI Cauchoix et al.

Author's name and affiliation:
Cauchoix Maxime[1,3], Hermer Ethan[2], Chaine Alexis S.[1,3], Morand-Ferron Julie[2]

[1]Station d'Ecologie Théorique et Expérimentale du CNRS UMR5321, Evolutionary Ecology Group, 2 route du CNRS, 09200 Moulis, France
[2]Department of Biology, University of Ottawa, Ottawa, Canada
[3]Institute for Advanced Studies in Toulouse, Toulouse School of Economics, 21 allée de Brienne, 31015 Toulouse, France

Author's contribution:
MC, JMF, ASC planned the study. MC and EH collected the data. MC analysed the data. MC, JMF and ASC wrote the SI methods.

Phil trans authors requested:
Ethan Hermer

Data sharing (full dataset or summary data):
Full datasets will be shared.

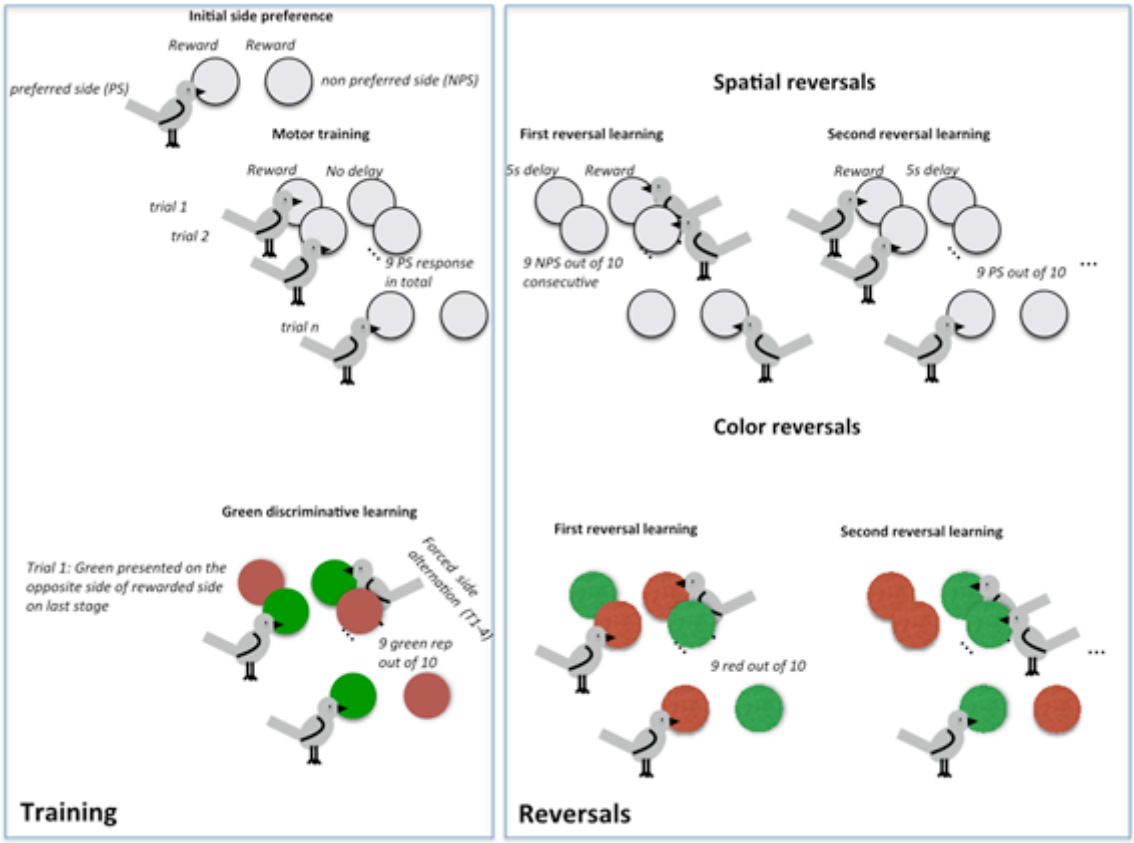


Figure1. Experimental procedure. Schematic description of training (left) and reversal (right) phases for serial spatial (top: Exp1, 2 and 3) and color (bottom: Exp3) reversals. Only performances in reversals are used for the repeatability analysis.

Methods. We collected data on spatial (Exp1, 2 and 3) and color (Exp 3) serial reversal learning performance of Great tits (Parus major) during 3 different experiments carried out in Moulis (France) in 3 different winter seasons:

- *Exp1: Cauchoix et al. wild,* from January to March 2015
- *Exp2*: *Cauchoix et al. group captivity*, from January to Mars 2016
- *Exp3: Cauchoix et al. isolated captivity*, from November 2016 to Mars 2017

For all experiments, data were collected using automated operant boxes that include two pecking keys and a reward hole as described in Morand-Ferron et al. (2015) using an individualized learning program for PIT-tagged birds (see Cauchoix et al. 2017 for more details). Experimental protocols are illustrated in Figure 1. In each experiment, birds were allowed one week of habituation to the operant boxes using dummy facades in which food was available at future key and reward-hole emplacements. Birds then began training for the cognitive tests with a side (left vs right) preference trial rewarded regardless of the side they pecked immediately followed by a motor training phase in which birds had to peck their initially preferred side 9 times (no punishment for pecks on the incorrect key). The first reversal began right after completion of this motor training phase. During the first reversal, the originally preferred key was not rewarded, but the non-preferred key was rewarded (Figure 1). In the spatial task, the two pecking keys were lit up in white and only one side was rewarded for each reversal while in the color task, one key was lit up in red and the other in green (random side) and only one color (regardless of side) was rewarded for each reversal. Birds had to reach a criterion of 9 correct responses out of 10 consecutive trials to complete a reversal and move to the following reversal in which the stimulus and reward contingencies were inversed. Cognitive performances were calculated as the accuracy (% of correct responses / total number of responses) and Trials To Criterion, or TTC (total number of trials before reaching the learning criterion of 9 correct response among 10 consecutives trials).

Experiment 1: wild

Twenty free-ranging great tits (12 male, 14 juvenile) performed from 1 to 48 serial spatial reversals in nature at two different study sites over 39 days. Each site was equipped with 2 operant boxes to minimize competition for access to the operant box. The methods and data are published in Cauchoix et al. (2017).

Experiment 2: group captivity

Seventeen wild-caught great tits (9 male, 15 juvenile) performed from 1 to 79 serial spatial reversals in captivity in groups of 2 to 5 individuals over 14 days. Each aviary 1 × 4 × 3 m (w × l × h) was equipped with 1 operant box. Housing conditions are described in Cauchoix et al. (2017). The methods and data are published in Cauchoix et al. (2017).

Experiment 3: isolated captivity

Twenty, individually housed, wild-captured great tits (9 male, 16 juvenile) performed from 1 to 20 serial spatial reversals in captivity (i.e. one bird per cage). Each aviary 1 × 4 × 3 m (w × l × h) was equipped with 1 operant box. All birds began with the same spatial reversal task described above and in Cauchoix et al (2017). After the 20th spatial reversal was completed, birds were given a colour discrimination learning task. Birds started with green as the rewarded stimuli because we had previously found a general initial preference for red in our great tit population (unpublished data). To aid the birds in switching their attention away from spatial cues to color cues, green was presented during the first color trial on the side opposite to the side previously rewarded in the 20th spatial reversal. We then forced color alternation between left and right for the 4 first trials and then assigned the side attributed to each color randomly. Trials were presented until birds pecked the rewarded color. The first reversal began after green discrimination learning was completed (i.e. green was pecked 9/10 trials over a sliding window)) and birds had to peck the red key to receive a reward. The whole experiment (spatial and color reversals) lasted for 14 days. Among the 20 birds, 9 birds completed between 1 and 10 color reversals. The data are unpublished.

References

Morand-Ferron, J., Hamblin, S., Cole, E. F., Aplin, L. M., & Quinn, J. L. (2015). Taking the operant paradigm into the field: associative learning in wild great tits. *PloS one*, *10*(8), e0133821.

Cauchoix, M., Hermer, E., Chaine, A. S., & Morand-Ferron, J. (2017). Cognition in the field: comparison of reversal learning performance in captive and wild passerines. *Scientific reports*, *7*(1), 12945.

8. SI Chow et al. 2015

Author's name and affiliation:
Chow PKY[1], Leaver LA[1], Wang M[2], Lea SEG[1]

[1]Centre for Research in Animal Behaviour, Department of Psychology, University of Exeter. United Kingdom. EX4 4QG
[2]Division of Biostatistics and Bioinformatics, Department of Public Health Sciences Penn State College of Medicine, Hershey, USA

Author's contribution:
PKYC and LA conceived and designed the experiments. PKYC performed the experiments. PKYC and WM analysed the data. PKYC contributed reagents/materials/analysis tools. PKYC and SEG wrote the paper.

Phil trans authors requested:
Pizza Ka Yee Chow

Data sharing (full dataset or summary data):
Data from the initial discrimination spatial learning phase and the first reversal spatial learning phase.

Methods:

- Five grey squirrels (Sciurus carolinensis) that were living in the Animal Cognition Laboratory participated the experiment.
- We tested 5 squirrels individually during their active period in the test room daily. They were not food- or water- deprived.
- We used a poke box to assess squirrels' spatial discrimination-reversal learning ability.
- The box was a square shaped box that had four wells at each corner.
- The base of the box was filled with hazelnut pieces and the whole box was wrapped by foil paper to control olfactory cues.
- The top of the box was covered by a white sheet.
- Squirrels were pre-trained to tear paper of a square poke box and obtained a hazelnut.
- Each well located at a concern. The task required squirrels to locate two (out of four) rewards that were baited at diagonal corner. Squirrels received four trial (max) each day and they had to make three consecutive trials correct (i.e choosing the rewarded well as 1st and 2nd choices). Once the squirrels had reached the learning criterion, we switched the reward contingency so that the previously unrewarded wells became rewarded whereas the previously rewarded wells became unrewarded.

Cognitive performance:

We measured the number of errors that each squirrel took to reach the learning criterion.

Reference

Chow PKY, Leaver LA, Wang M, Lea SEG. (2015). Serial reversal learning in grey squirrels: learning efficiency as a function of learning and change of tactics. Journal of Experimental Psychology: Animal Learning and Cognition, 41, 343-353. http://psycnet.apa.org/record/2015-30042-001

9. SI Chow et al. 2016

Author's name and affiliation:
Chow PKY, Lea SEG, Leaver LA

Centre for Research in Animal Behaviour, Department of Psychology, University of Exeter. United Kingdom. EX4 4QG

Author's contribution:
PKYC designed and conducted the experiments. PKYC analysed the data and wrote the first draft of the paper. All authors revised the paper.

Phil trans authors requested:
Pizza Ka Yee Chow

Data sharing (full dataset or summary data):
First trial of the data is shared

Methods:

- Five grey squirrels (Sciurus carolinensis) that were living in the Animal Cognition Laboratory participated the experiment.
- We tested 5 squirrels individually during their active period in the test room. They were not food- or water- deprived.
- We assessed squirrels' problem-solving ability by giving squirrels a puzzle box. This puzzle box had 10 levers (5 with hazelnuts and 5 without hazelnuts). Squirrels had to use alternative methods to make a lever/nut dropped to obtain a success.
- Before the experiment, we included a habitation phase for squirrels to minimise their neophobic response toward novel stimulus.
- The whole experiment contained three blocks of four trials. There was a one-day break between blocks (total 14 days).

Cognitive performance
We measured solving duration, considered as when a squirrel started to use any of their body parts to manipulate a lever until it made a lever/nut dropped.

Reference
Chow PKY, Lea SEG, Leaver LA (2016). How practice makes perfect: the role of persistence, flexibility and learning in problem-solving efficiency. Animal Behaviour, 112, 273-283. https://doi.org/10.1016/j.anbehav.2015.11.014

10. SI Chow et al. 2017a

Author's name and affiliation:
Chow PKY, Lea SEG, Hempel de Ibarra N, Robert T

Centre for Research in Animal Behaviour, Department of Psychology, University of Exeter. United Kingdom. EX4 4QG

Author's contribution:
PKYC and TR designed the experiment. PKYC conducted the experiments, analysed the data and wrote the first draft of the paper. All authors revised the paper.

Phil trans authors requested:
Pizza Ka Yee Chow

Data sharing (full dataset or summary data):
First trial of the data is shared

Methods:

- The same five grey squirrels (Sciurus carolinensis) that had participated the puzzle box experiment participated this experiment.
- We tested them individually during their active period in the test room. They were not food- or water- deprived.
- We assessed squirrels' memory for the successful solution in solving the puzzle box. To do so, we gave squirrels the same box 22 months after their last experience with it (recall task).
- 6 days after the recall task, we then assessed squirrels' ability to generalise the successful solution to a similar but different box (triangle box). This triangle box had 5 levers and each lever had one hazelnut. Squirrels had to apply the same successful solution to obtain a hazelnut as success.
- Both experiments had the same procedures as with the last experiment in Chow et al., 2016; each experiment had three blocks of four trials. There was a one-day break between blocks (total 14 days).

Cognitive performance
We measured solving duration, considered as when a squirrel started to use any of their body parts to manipulate a lever until it made a lever/nut dropped.

Reference
Chow PKY, Lea SEG, Hempel de Ibarra N, Robert T. (2017). How to stay perfect: the role of memory and behavioural traits in an experienced problem and a similar problem. Animal Cognition, 20, 941-952. https://doi.org/10.1007/s10071-017-1113-7

11. SI Chow et al. 2017b

Author's name and affiliation:
Chow PKY[1], Leaver LA[1], Wang M[2], Lea SEG[1]

[1]Centre for Research in Animal Behaviour, Department of Psychology, University of Exeter. United Kingdom. EX4 4QG
[2] Division of Biostatistics and Bioinformatics, Department of Public Health Sciences Penn State College of Medicine, Hershey, USA

Author's contribution:
PKYC and LAL designed the experiments. PKYC performed the experiments and PKYC and WM analysed the data. PKYC wrote the first draft of the paper and all authors revised the paper.

Phil trans authors requested:
Pizza Ka Yee Chow

Data sharing (full dataset or summary data):
Full dataset is shared

Methods:

- Five grey squirrels (Sciurus carolinensis) that were living in the Animal Cognition Laboratory participated the experiment.
- We tested 5 squirrels individually during their active period in the test room daily. They were not food- or water- deprived.
- We setup a touchscreen for this experiment in a test room that squirrels were familiar with.
- Squirrels were pre-trained to use the touchscreen before the main experiment started.
- Before the main testing, we tested squirrels' colour preference by presenting two colours stimuli, one green and one red, to them. Colour preference was indicated as choosing 3 consecutive colour out of five trials. Less preferred food reward was given at this preference test.
- Four squirrels showed a preference to green (G+ R-) and hence we rewarded their non-preferred colour in the discrimination learning phase whereas one squirrels (Sarah) did not showed any preference to either green or red colour and hence we first trained Sarah to show a preference to green before her main experiment started.
- Squirrels received a block (60 trials) per day.
- Once they reached a learning criterion, we switched the reward contingency (G- R+) for squirrels in the reversal learning phase, and squirrels had to learn the new contingency until they reached the learning criterion again.

Cognitive performance
For each phase, we measured the number of errors that a squirrel made until they reached the

learning criterion 75% for two consecutive blocks for four squirrels and 70% for one squirrel, Simon.

Reference

Chow PKY, Leaver LA, Wang M, Lea SEG. (2017). Touchscreen assays of behavioural flexibility and error characteristics in Eastern grey squirrels (Sciurus carolinensis). Animal Cognition, 20, 459-471. https://link.springer.com/article/10.1007%2Fs10071-017-1072-z

12. SI Chow (unpublished data)

Author's name and affiliation:
Pizza Ka Yee Chow, Heather Honan
Centre for Research in Animal Behaviour, Department of Psychology, University of Exeter. United Kingdom. EX4 4QG

Author's contribution:
PKYC designed the experiments. PKYC and HH performed the experiments. HH analysed the data.

Phil trans authors requested:
Pizza Ka Yee Chow

Data sharing (full dataset or summary data):
Full dataset is shared

Methods:

- Data were collected between Dec, 2015 and Jan, 2016 with three squirrels that had participated the colour discrimination-reversal learning task on touchscreen.
- All data were collected when squirrels were active in their home cage.
- They were not food- or water- deprived during the experiment.
- Before the main experiment, squirrels went through a habitation phase; this aimed to minimize their neophobic responses toward the mesh tube.
- We used five cylinder mesh tubes (12cm Length x 7.5 diameter), with the middle of the tube was baited with a hazelnut.
- One side of the tube was covered by a black colour plastic disk and the other was a white colour plastic disk**.**
- The five mesh tubes were placed as two parallel rows in a large test cage.
- Similar to the touch screen study, we tested squirrels' colour preference with five trial and a preference for either colour was indicated as 3 (out of five) consecutive trials.
- None of the squirrels showed a preference to either colour and thus the first rewarded colour was counterbalanced across squirrels in the discrimination learning phase.
- We gave squirrels one trial per day and squirrels had to choose 4 (out of five) rewarded colour as first choice for two consecutive days in order to past the learning criterion.
- Once they reached the learning criterion, we switched the reward contingency so that the previously rewarded colour became unrewarded whereas the previously unrewarded colour became rewarded.
- Squirrels had to learn the new reinformance contingency until they reached the learning criterion again.

Cognitive performance
We measured the number of error squirrels made until they reached the learning criterion.

13. SI Chow et al. 2018

Author's name and affiliation:
Chow PKY[1], Lurz PWW[2], Lea SEG[1]

[1]Centre for Research in Animal Behaviour, Department of Psychology, University of Exeter. United Kingdom. EX4 4QG
[2]The Royal (Dick) School of Veterinary Studies, Easter Bush Campus, University of Edinburgh, Midlothian EH25 9RG

Author's contribution:
PKYC designed the experiments, performed the experiments and analysed all behavioural data. PKYC wrote the first draft and all authors revised the paper.

Phil trans authors requested:
Pizza Ka Yee Chow

Data sharing (full dataset or summary data):
Figures, solving duration on the first success in the easy motor task and the difficult motor task will be shared.

Methods. We collected data on solving food-extraction problems in Eurasian red squirrels and Eastern grey squirrels in the field.

Experimental setup. This was a field experiment carried out between Sept and Nov, 2014 for the red squirrels on Isle of Arran, Scotland and between Dec, 2014 and Feb, 2015 for grey squirrels in Exeter, England. We collected data at seven locations in woodland and eight locations in woodland around Exeter campus or in the campus itself. Both tasks are set in bushes or away from major road. Squirrels were individually identified by their body shape and colour, tail and ear shape. Tasks were presented to squirrels from dawn to dusk. We also counterbalanced the presentation of tasks. We checked each task every 1-2 hours.

Cognitive test
An easy task: 17 red squirrels and 14 grey squirrels participated the task.
A difficult task: 13 red squirrels and 20 grey squirrels participated the task.

The easy task was a hinged box (Figure 1a) Squirrels were required to lift up a lid to obtain a hazelnut in the easy task. The difficult task was a puzzle box that were reported in Chow et al., 2016 and Chow et al., 2017 (Figure 1b). It was a transparent Plexiglas box. Each side of the box had 10 holes that are horizontally but not vertically aligned. Each hole has a lever which squirrels had to act on. Each lever had a three-sided nut container. After inserting a lever into the box, 2.5 cm of lever end protruded outside of the box. Squirrels had to push a lever on the lever end that was close to the nut container. to obtain a hazelnut in the difficult task.

Figure 1. a) the easy task (left): a hinged box that had four well with each well attached with a transparent lid. Squirrels had to lift up a lid to obtain a hazelnut. b) the difficult task (right): a transparent Plexiglas box. Each side of the box has 10 holes that are horizontally but not vertically aligned. Each lever had a three-sided container located on one end of a lever. Squirrels had to push in a lever (if they are close to the nut container) or pull a lever (if they are far from the nut container) to obtain a hazelnut.

Cognitive performance

Problem-solving performance was measured as solving outcome (success or failure) at a squirrel's first visit (indicated as first appeared on the video until it left the video for 2 or more minutes) or subsequent visit (indicated as the same squirrels re-appeared the on the video). We also measured solving duration (indicated as when a squirrels first manipulated an apparatus using any of its body part until it stopped the contact). This solving duration incorporated all unsuccessful solving duration until a success occurred.

Reference

Chow PKY, Lurz PWW, Lea SEG (2018). A battle of wits? Problem-solving abilities in invasive Eastern grey squirrels and native Eurasian red squirrels. *Animal Behaviour,* 137, 11-20.

14. SI Claidière et al.

Author's name and affiliation:
Nicolas Claidière[1], Gameli Kodjo-kuma Amedon[1], Jean-Baptiste André[2], Simon Kirby[3], Kenny Smith[3], Dan Sperber[4,5], Joël Fagot[1]
[1]Aix Marseille Univ, CNRS, LPC, Marseille, France
[2]Institut des Sciences de l'Evolution, CNRS, Montpellier, France
[3]Centre for Language Evolution, School of Philosophy, Psychology, and Language Sciences, University of Edinburgh, Edinburgh EH8 9AD, United Kingdom
[4]Departments of Cognitive Science and of Philosophy, Central European University, Budapest, Hungary
[5]Institut Jean Nicod (CNRS, EHESS, ENS), Paris, France

Phil trans authors requested:
Nicolas Claidière, Joël Fagot

Data sharing (full dataset or summary data):
Data are available in the Open Science Framework repository:
https://osf.io/h84y6/
DOI 10.17605/OSF.IO/H84Y6

Methods. Guinea baboons (Papio papio) belonging to a large social group of the CNRS Primate Center in Rousset-sur-Arc (France) participated in this study. The baboons were all marked by two biocompatible 1.2 by 0.2 cm RFID microchips injected into each forearm and lived in an outdoor enclosure (700m2) connected to an indoor area which provided shelter when necessary. Baboons were neither water- nor food-deprived during the research. Water was provided ad-libitum within the enclosure. Monkeys received their normal ration of food (fruits, vegetables and monkey chows) every day around 5 pm. The baboons were all born within the primate centre.

Self-testing procedure
Experiments were conducted in the facility developed by Joël Fagot. The key feature of this facility is that baboons have free access to computerized testing booths that are installed in trailers next to their enclosure. They can thus participate in experiments whenever they choose, and do not need to be captured (more details can be found in Fagot & Bonté, 2010; Fagot, Marzouki, Huguet, Gullstrand, & Claidière, 2015; Fagot & Paleressompoulle, 2009). The voluntary participation of the subjects reduces stress levels, as inferred from the significant decrease in salivary cortisol levels as well as the frequency of stereotypies (Fagot, Gullstrand, Kemp, Defilles, & Mekaouche, 2014).

Computer-based tasks
During the experiment, each computerised trial began with the display of a grid made of 16 squares, 12 white and 4 red (Claidière, Smith, et al., 2014). Touching this stimulus display triggered the immediate abortion of the trial and the display of a green screen for 3 seconds

(time-out). After 400 ms all the red squares became white and, to obtain a food reward, the monkey had to touch the previously red squares, in any order and with less than 5 seconds between touches. Squares became black when touched to avoid being touched again and did not respond to subsequent touches. The trial was completed when 4 different squares had been touched. If three or four correct squares were touched the trial was considered a success and the computer triggered the delivery of a reward. If less than 3 correct squares were touched the trial was considered a failure and a green time out screen appeared for 3 seconds.

Training

All members of the colony underwent a training procedure to enable them to participate in the experiment: only those animals who reached our final criterion were admitted into the experiments. Training followed a progressive increase in the complexity of the task, starting with only one target (red square), followed by a stage with one target and one distractor (white square), then by an increase in targets up to four and finally by an increase in the number of distractors up to 12. Training blocks consisted of 50 trials and progress through training was conditioned on performing above criteria (80% success on a block of 50 random trials, excluding aborted trials (on average 1.7% (SD=0.98%) of trials), which were re-presented).

Testing

In this study we analyze the results of the baboons when they were presented with a succession of randomly selected grids among the set of all possible grids. These trials, always following exactly the same procedure, were performed at three different times, during October and November 2012, January 2013 and January and February 2014.

Ethics statement

This research was carried out in accordance with French and EU standards and received approval from the French Ministère de l'Education Nationale et de la Recherche (approval # APAFIS-2717-2015111708173794-V3). Procedures were also consistent with the guidelines of the Association for the Study of Animal Behaviour.

References

Claidière, N., Smith, K., Kirby, S., & Fagot, J. (2014). Cultural evolution of systematically structured behaviour in a non-human primate. Proceedings of the Royal Society B: Biological Sciences, 281(1797).

Claidière, N., et al. (2018). Convergent transformation and selection in cultural evolution. Evolution and Human Behavior.

Fagot, J., & Bonté, E. (2010). Automated testing of cognitive performance in monkeys: Use of a battery of computerized test systems by a troop of semi-free-ranging baboons (Papio papio). Behavior research methods, 42(2), 507-516.

Fagot, J., Gullstrand, J., Kemp, C., Defilles, C., & Mekaouche, M. (2014). Effects of freely accessible computerized test systems on the spontaneous behaviors and stress level of Guinea

baboons (Papio papio). American Journal of Primatology, 76(1), 56-64.

Fagot, J., Marzouki, Y., Huguet, P., Gullstrand, J., & Claidière, N. (2015). Assessment of Social Cognition in Non-human Primates Using a Network of Computerized Automated Learning Device (ALDM) Test Systems. JoVE(99), e52798.

Fagot, J., & Paleressompoulle, D. (2009). Automatic testing of cognitive performance in baboons maintained in social groups. Behavior research methods, 41(2), 396-404.

15. SI Dalesman 2015

Author's name and affiliation:
Dalesman, Sarah[1]

[1]Institute of Biological, Environmental and Rural Sciences, Aberystwyth University, Penglais Aberystwyth, Ceredigion SY23 3DA. U.K.

Author's contribution:
SD collected the data and wrote the methods.

Phil trans authors requested:
Sarah Dalesman

Data sharing (full dataset or summary data):
Full datasets will be shared.

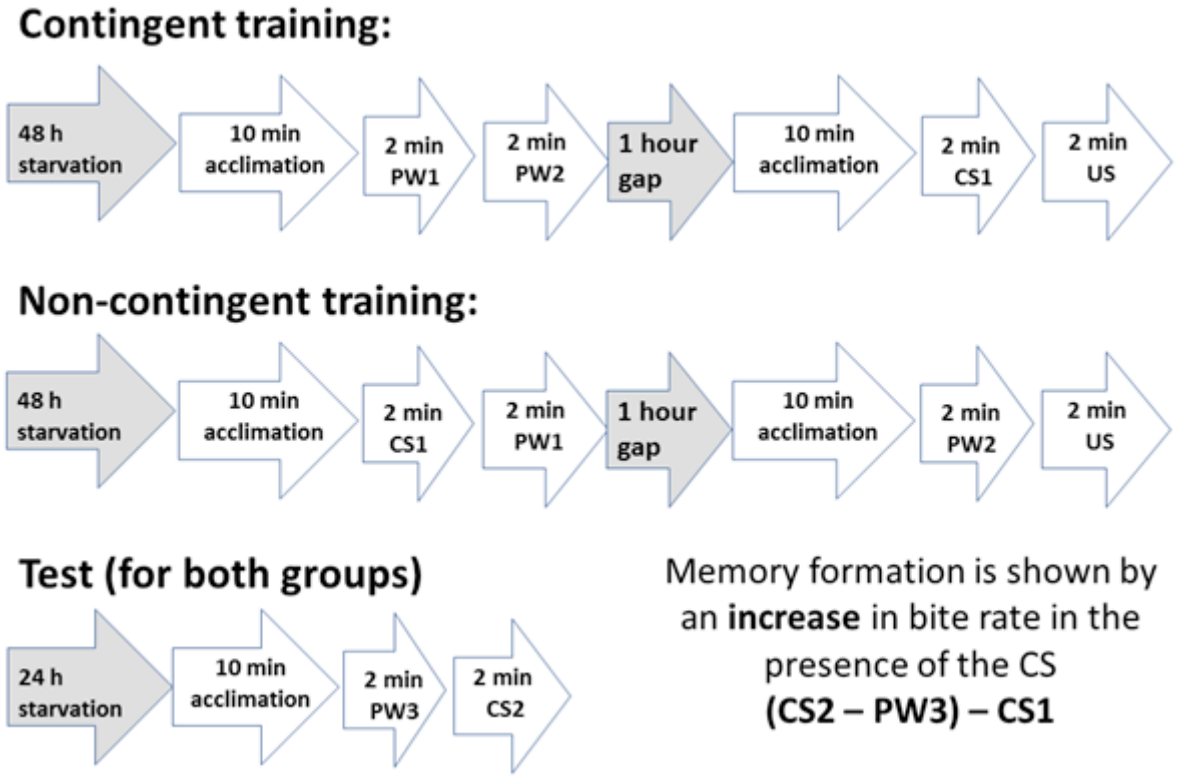


Figure 1. Experimental procedure for appetitive food conditioning. Schematic description of contingent training (a) and non-contingent training (b). CS (conditioned stimulus) = amyl acetate or gamma-nonalactone (both 0.004 %), US (unconditioned stimulus) = sucrose solution (0.67 %), PW = artificial pond water. Bite rate to determine memory formation is recorded during CS1, PW3 and CS2. Only performances in contingently trained animals are used for the repeatability analysis.

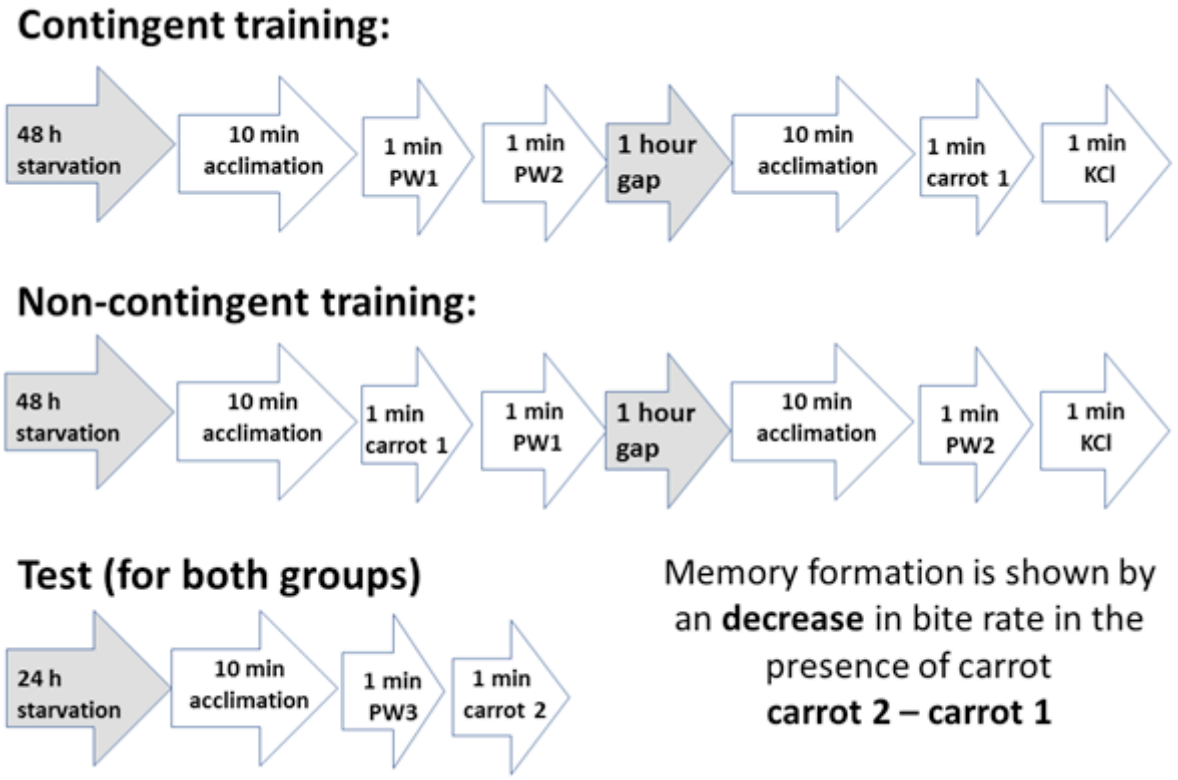


Figure 2. Experimental procedure for aversive food conditioning. Schematic description of contingent training (a) and non-contingent training (b). Carrot = carrot solution (0.25 %), KCl = potassium chloride solution (5 mM), PW = artificial pond water. Bite rate to determine memory formation is recorded during carrot 1 and carrot 2. Only performances in contingently trained animals are used for the repeatability analysis.

Methods. I collected data on appetitive food conditioning (Expt. 1 & 2), aversive food (Expt. 2) and conditioning operant conditioning (Expt. 3) in pond snails, *Lymnaea stagnalis*, in three separate experiments.

Experiment 1. Comparing appetitive and aversive food conditioning – part of a data set published in Dalesman *et al.* (2015), collected June to September 2013 at the University of Exeter.
Experiment 2. Repeated appetitive conditioning using 2 different conditioned stimuli, collected October to November 2017 at Aberystwyth University.
Experiment 3. Repeated operant conditioning, collected June to August 2016 at Aberystwyth University

All experiments were carried out using F1 adult *Lymnaea stagnalis* reared in the laboratory from wild collected adults. Experiment 1 used adults from four different river sites, South Drain (R1), Sowy River (R2) and Main Drain (R3) on the Somerset Levels and Exeter Canal (R4) (see Dalesman *et al.* 2015). Experiments 2 and 3 used snails from South Drain only. Snails were maintained in grouped conditions throughout at a density of approximately one snail per 0.5 l in 10 l aquaria in artificial pond water (see Dalesman *et al.* 2015 for details) on a

14:10 L:D cycle at 20^{o}C. 48 hours prior to starting each experiment snails were individually labelled using bee tags (E.H. Thorne (Beehives) Ltd.) attached using Loctite 454 (Henkel). Unless food deprived as part of an experimental trial snails were fed *ad libitum* on round lettuce and trout pellets.

Experiment 1

Eighty snails from four river populations were tested across three different memory tasks, operant conditioning of aerial respiration, aversive food conditioning and appetitive food conditioning. The order in which each task was performed was allocated in a randomized block design (see Dalesman *et al.* 2015). Snails were either trained using operant (blocks 1 and 3) or food conditioning (blocks 2 and 4) first, and within the food conditioning trials they either received aversive food conditioning (1 and 2) or appetitive conditioning (3 and 4) first. Eighty snails were used as non-contingent controls, with training carried out alongside all contingently trained snails. Non-contingent control snails received identical stimuli over the course of the experiment (Figure 1 & 2), but the CS and US were not contingent. Non-contingent training did not result in a change in behaviour between training and testing (Dalesman *et al.* 2015). Data from contingently trained snails in the aversive and appetitive conditioning tasks are analysed here for repeatability, using the change in bite rate between training and testing to determine long-term memory formation (Figure 1 & 2).

The methods and full data are published in Dalesman *et al.* (2015).

Experiment 2

Forty snails were trained using contingent appetitive food conditioning, and forty snails were trained as non-contingent controls (Figure 1). Snails were trained in two blocks, block 1 starting 21st October 2017, block 2 starting 11th November 2017 with 20 contingent and 20 non-contingently trained snails per block (see below). In each block snails received two training trials separated by 2 weeks to allow recovery from food deprivation prior to the subsequent trial. Half the snails were trained using amyl acetate (Marra *et al.* 2013) as the conditioned stimulus and half using gamma-nonalactone (Ildiko Kemenes, pers. comm.) as the conditioned stimulus in trial 1. In trial 2 they were then swopped such that snails which were initially trained with amyl acetate were then trained with gamma-nonalactone, and *vice versa*. Snails do not generally respond to either amyl acetate or gamma-nonalactone as a food resource, and demonstrated no change in behaviour following non-contingent training. Sucrose, which causes a strong increase in feeding behaviour in food deprived snails, was used as the unconditioned stimulus throughout.

Contingent training (Figure 1): Following 48 h food deprivation in home aquaria, snails were placed individually in a small Petri dish (55 mm diameter x 12 mm height) above a mirror in 18 ml of pond water to acclimate for 10 min. 1 ml of pond water we then added (PW1), followed 2 min later by a further 1 ml of pond water (PW2) and left for a further 2 min before returning to their home aquaria. One hour later they received a further training session, 10 min acclimation in a Petri dish in 18 ml of pond water. During the second session 1 ml of the CS (CS1: 0.08 % solution of either AA or GN giving a final concentration of 0.004 %; Figure 1)

was then added, followed 2 min later by addition of 1 ml of the US (13.4 % sucrose solution – giving a final concentration of 0.67%; Figure 1) and left for a further 2 min before returning to their home aquaria. Twenty-four hours later snails were tested for their response to the conditioned stimulus. During the test snails were again acclimated in the Petri dish in 18 ml of pond water for 10 min, following which 1 ml of pond water (PW3) was added to the Petri dish, followed 2 min later by additions of 1 ml of the CS they had experienced the previous day (CS2). Bite rate was recorded during CS1, PW3 and CS2 by viewing the snails behaviour in the mirror (Figure 3).

Non-contingent training: This procedure is identical to the procedure described above, except that the conditioned stimulus (CS1) is swopped with the first pond water exposure (Figure 1). Therefore, the snails receive an identical handling experience and identical time exposed to each of the stimuli; however, the CS and US are not experienced contingently during training. If associative memory is responsible for the change in bite rate during the test, rather than a general sensitisation to the CS on repeated exposure, then non-contingently trained snails should not demonstrate a change in bite rate in response to the CS.

Long-term memory is measured as a change in bite rate in response to the CS between training and the test 24 h later. This value is adjusted for the bite response in pond water alone during the test period (Figure 1: PW3) to account for any change in behaviour that is not due to the response to the conditioned stimulus.

Only data from contingently trained animals has been used to determine repeatability. The data are not previously published.

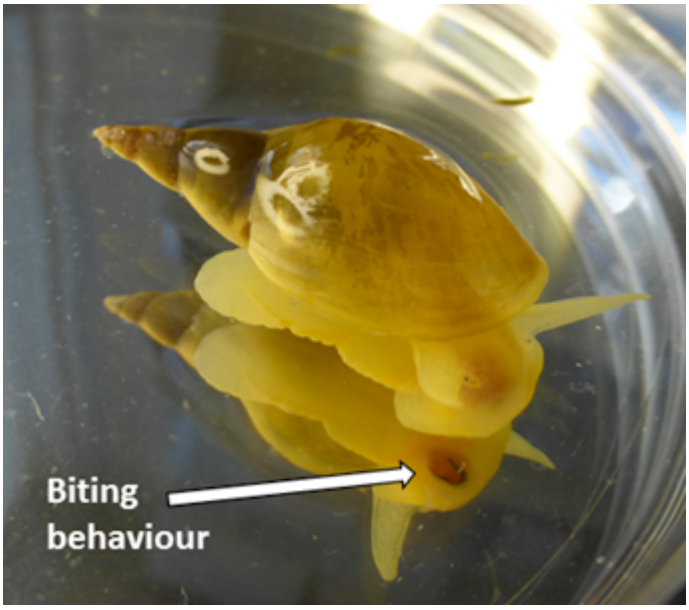


Figure 3. Biting behaviour recorded during appetitive and aversive conditioning. The rasping behaviour is clearly visible when viewed over a mirror. Image: S. Dalesman

Experiment 3

Thirty-six snails were trained in using contingent training and thirty-six using non-contingent training of operant conditioning of aerial respiration (Lukowiak *et al.* 1996) between June and August 2016 in three blocks started one week apart. One snail was excluded from contingent training as it died prior to the final training session, and its match in the non-contingent group is also excluded from the data. Each individual snail within a block received three trials, trial 1 and 2 were separated by 1 week, trial 2 and 3 were separated by 4 weeks. Data from contingent training only has been used to determine repeatability.

Contingent training: N_2 was vigorously bubbled through 600 ml of artificial pond water in a 1 l beaker for 20 mins to make the water hypoxic (approx. 5% O_2 saturation), at which point the bubbling was reduced to prevent disturbance to the snails during the training trial. Snails were introduced to the beaker in groups of 6 individuals and allowed to acclimate for 10 min. During the 30 min training snails were gently poked on the pneumostome each time they came to the surface of the water and attempted to breathe using a wooden skewer (Figure 4). The poke is sufficient that it causes the snail to close the pneumostome, but does not cause full body withdrawal. The number of pokes an individual received during the training trial was recorded. Snails were returned to their home aquaria between training sessions. Twenty-four hours following training snails were tested for long-term memory formation using an identical procedure. Memory formation is determined by the change in breathing attempts between training and testing, a reduction in breathing attempts is considered to demonstrate memory formation.

Non-contingent training: At the same time as snails are trained contingently, a second beaker is prepared in an identical manner for the non-contingent control training (yoked controls). Each snail in the yoked control group is randomly paired with a snail in the contingently trained group. During training, the snail in the yoked control group is then poked in the vicinity of the pneumostome (or on the pneumostome if it happens to be open at the time) when it's partner is poked contingently with pneumostome opening. Therefore, the yoked animals received an identical number of stimuli during training to the contingently trained individuals. During the test phase 24 hours later, the yoked control snails receive a poke contingent with pneumostome opening. A lack of change in breathing attempts in the yoked controls allows me to determine that contingency is required for memory formation, rather than a generalized response to hypoxia or physical stimulation.

Only data from contingently trained animals has been used to determine repeatability. The data are not previously published.

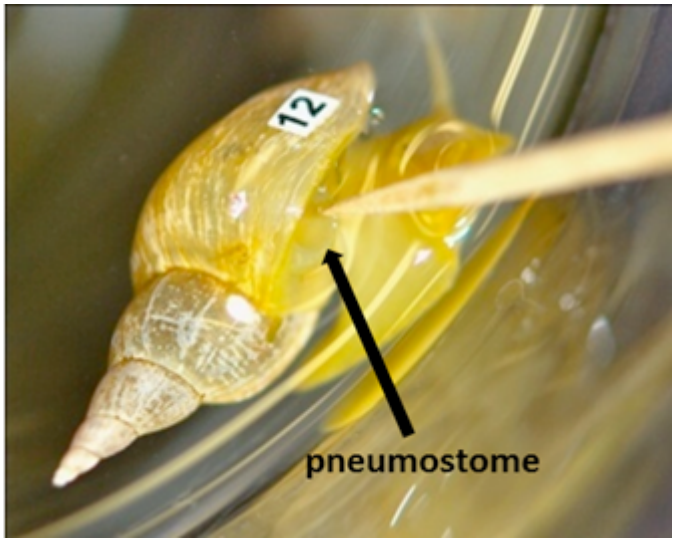


Figure 4. *Lymnaea stagnalis* undergoing operant conditioning of aerial respiration. The pneumostome has just opened and the snail is receiving a poke with a wooden skewer. Image: S. Dalesman

References

Dalesman, S., Rendle, A., & Dall, S. R. X. (2015). Habitat stability, predation risk and 'memory syndromes'. Scientific reports, 5, 10538. doi: 10.1038/srep10538

Lukowiak, K., Ringseis, E., Spencer, G., Wildering, W., & Syed, N. (1996). Operant conditioning of aerial respiratory behaviour in *Lymnaea stagnalis*. Journal of Experimental Biology, 199(3), 683-691.

Marra, V., O'Shea, M., Benjamin, P. R., & Kemenes, I. (2013). Susceptibility of memory consolidation during lapses in recall. Nature Communications, 4, 1578. doi: 10.1038/ncomms2591

16. SI Henke-von der Malsburg & Fichtel 2015

Author's name and affiliation:
Henke-von der Malsburg Johanna[1], Fichtel Claudia [1]

[1]Behavioural Ecology and Sociobiology Unit, German Primate Centre, Kellnerweg 4, 37077 Göttingen

Author's contribution:
JHM and CF conceived the study and developed the experimental design. First experiments were conducted by JHM and CF, whereas JHM conducted remaining experiments and analysed the video-recordings.

Phil trans authors requested:
Johanna Henke-von der Malsburg, Claudia Fichtel

Data sharing:
Summary data will be shared.

Methods. From September to October 2015 we tested 7 wild Grey mouse lemurs (*Microcebus murinus*; 60 g) and 1 wild Madame Berthe's mouse lemur (*Microcebus berthae*; 30 g) living in sympatry in Kirindy Forest, Western Madagascar, on their discriminative and subsequent reversal learning abilities. Animals were captured with Sherman life traps and housed individually in cages of about 80 x 80 x 80 cm^3 in the research station for a maximum of three consecutive nights. If an individual did not finish the experiments within these three nights, testing was continued as soon s as the individual was recaptured. Experiments were conducted during night under dimmed red-light conditions and video-taped (Sony HDR-CX 240).

We conducted two discrimination learning and subsequent reversal tasks addressing different sensory cues.

- Visual discrimination: use of two differently shaped small plastic forms (octagonal- and semicircle-shaped cylinder)
- Olfactory discrimination: use of two different solvents (water, peppermint syrup) put into a lid placed within a funnel to bundle the odour in one direction

We randomly differed the first discrimination modality between the individuals and conducted the respective second discrimination modality after the reversal task of the first discrimination modality was finished.

Prior to the actual experiment we trained the animals to indicate their choice by reaching for a plastic form or solvent, respectively, placed on a presentation board outside of their cage from inside their cage. The plastic form and solvent that were used for the training were different from the one used during testing.

For the discrimination learning task, we conducted sessions of 11 trials where only one stimulus was rewarded, which one, was previously chosen and differed randomly between the subjects. A trial started when the presentation board with the two stimuli was placed in front of the cage and ended with the individual having reached for one of the two stimuli or after a maximum of 60 seconds (refusal). The session was stopped earlier if the subject refused to

choose for three consecutive trials. The positions of the stimuli differed randomly within a session to avoid side preferences. Once the low-level learning criterion of at least 7 correct choices within a given session (67% correct choices; Rumbaugh & Gill 1972) was reached, we reversed the rewarding stimulus for the reversal task. The reversal task consisted of a single session with 11 trials where the first trial served as signal to the reversed rewarding and only the subsequent 10 trials were analysed.

We counted the trials needed to reach the learning criterion in the discrimination learning task and the percentage of correct choices in the last (criterion) session (11 trials) of the discrimination learning task and in the single session of the reversal task (10 trials).

Ethical note

This study was conducted in accordance with the German and Malagasy (Commission Tripartite CAFF) legal and ethical requirements of appropriate animal procedures. Research protocols and experimental procedures were approved by the Ministry for the Environment, Water and Forests of Madagascar (MINEEF).

Reference

Rumbaugh DM, Gill T V. (1973) The learning skills of great apes. J Hum Evol 2:171–179. doi: 10.1016/0047-2484(73)90073-0

17. SI Henke-von der Malsburg & Fichtel 2016

Author's name and affiliation:
Henke-von der Malsburg Johanna, Fichtel Claudia

Behavioural Ecology and Sociobiology Unit, German Primate Centre, Kellnerweg 4, 37077 Göttingen

Author's contribution:
JHM and CF conceived the study and developed the experimental design. First experiments were conducted by JHM and CF, whereas JHM conducted remaining experiments and analysed the video-recordings.

Phil trans authors requested:
Johanna Henke-von der Malsburg, Claudia Fichtel

Data sharing:
Summary data will be shared.

Methods. From August to October 2016 we tested 23 wild Grey mouse lemurs (Microcebus murinus; 60 g) and 12 wild Madame Berthe's mouse lemurs (Microcebus berthae; 30 g) living in sympatry in Kirindy Forest, Western Madagascar, in a three-task problem-solving experiment using an artificial wooden box (11.5 x 7.5 x 3.0 cm^3). Animals were captured with Sherman life traps and housed individually in cages of about 80 x 80 x 80 cm^3 in the research station for a maximum of three consecutive nights. If an individual did not finish the experiments within these three nights, testing was continued as soon s as the individual was recaptured. Experiments were conducted during night under dimmed red-light conditions and video-taped (Sony HDR-CX 240).

The artificial box had a hinged door on each side which could be blocked in either a closed or an opened state. Additionally, a drawer could be inserted in the box. We varied the opening possibilities of the box to create two novel problems and one modified problem in which the animals should retrieve a food reward (piece of banana) from inside the box and which we presented one after another (Figure 1).

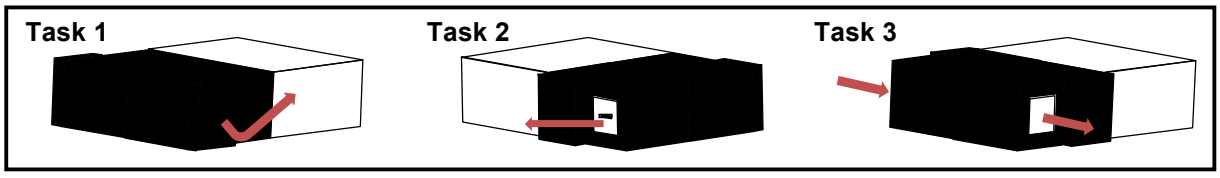


Figure. 1 Closed stages of the novel problem-solving box in the three tasks. Arrows indicate the direction of the opening mechanism. Striped parts were painted light blue, dotted parts were painted dark brown.

In the first novel problem only one hinged door could be opened to retrieve the reward. This door was blocked in the second novel problem where instead, the respective other door was opened and the drawer containing the food reward was inserted inside the box. The animals could pull out the drawer to retrieve the reward. The last task was a modification of the second novel problem, where the drawer was also inserted in the box, but the possibility to pull it out was blocked and both doors were opened. The animals, thus, had to push the drawer out of the box and change the sides to retrieve the reward.

Before the actual experiment, we conducted a familiarization phase, where we put the box, baited with a piece of banana on each side, with both doors opened and without drawer, onto the experimental platform within the subject's cage. The animals had to retrieve the rewards six times on each side to ensure that they have learned to retrieve the reward from both sides of the box.

For the experiment, we conducted sessions consisting of 12 trials. A trial started when the box was placed onto the experimental platform and stopped either when the animal had retrieved the reward or when five minutes passed by (10 minutes in the first trial of the first novel problem). A session was stopped earlier if the subject did not inspect the box in two trials. Before each session, we cleaned the box with 80% ethanol to exclude olfactory cues left by previous subjects that may guide the next one to the solution. We moved on to the respective following task when the subject reached the learning criterion of retrieving the reward in at least 10 trials during a given session.

Video-recordings were analysed at half speed by using Windows Movie Maker version 16.4.3528.0331 (© 2012 Microsoft Corporation). We recorded the success latency as time from the individual's approach until its success (retrieving the reward), the number of errors, i.e. the number of unsuccessful trials prior to the first successful trial, the number of successful and unsuccessful trials, and subsequently the success rate as proportion of successful trials to total trials. For the success latency, we calculated the arithmetic mean for each individual per task. A second person naïve to the research question analysed 10% of the videos a second time to assess interrater reliability, which was 99.3% (Intraclass correlation coefficient; R package 'ICC', Wolak 2015).

Ethical note

This study was conducted in accordance with the German and Malagasy (Commission Tripartite CAFF) legal and ethical requirements of appropriate animal procedures. Research protocols and experimental procedures were approved by the Ministry for the Environment, Water and Forests of Madagascar (MINEEF).

Reference

Wolak M (2015) Facilitating estimation of the Intraclass Correlation Coefficient. ICC Package 1–9.

18. SI Huebner & Kappeler 2018

Author's name and affiliation:
Franziska Huebner[1], Peter M. Kappeler [1,2]

[1]Behavioral Ecology & Sociobiology Unit, German Primate Center, Germany, [2] Department of Sociobiology/ Anthropology, University of Göttingen, Germany

Author's contribution:
FH, PMK planned the study. FH collected and analysed the data. FH wrote the SI methods.

Data sharing:
Partly the full dataset and partly a summary will be shared as the data are still unpublished.

Methods

We collected data on two novel problem-solving tasks, a food extraction task and a string-pulling task, with wild grey mouse lemurs (*Microcebus murinus*) in Kirindy forest, Western Madagascar. Data were collected in three consecutive dry seasons between April and July in 2015- 2017. For the testing, mouse lemurs were captured with Sherman live traps and housed in cages ($65cm^3$) for up to three consecutive nights in the research station. Water was provided *ad libitum*, and food was offered directly after the cognitive testing. Animals were tested in their cages at night under dim red light and each test session was video-recorded. After the short-term captivity, individuals were released to their specific site of capture and if possible recaptured after 10 to 30 days and/ or about one year for the repeatability tests (for more information on the general procedure see Huebner et al. 2018 for more details). We tested a total of 96 juvenile (of about 3-7 months of age) and adult subjects of both sexes in both tasks, with varying sample sizes for the repetitions (see Table 1).

Food extraction task

The food-extraction task consisted of removing a sliding cover on each of the 6 wells (5 x 4.5cm) of a small box (6 x 12cm) in order to access a little banana reward in each compartment (Figure 1). Before actual testing, subjects were attracted to this task with a small, freely accessible banana piece on top of the box, more habituation was not needed. Subjects were presented with the novel problem for a maximum of 20 minutes during which we recorded whether subjects were able to open at least one lid (success y/n), the number of total successes (0 to 6) and their latency from first contact with the box to first success. For subjects that did not succeed, we noted their total duration of testing, starting with the first contact with the box (i.e., capped latencies). Further, we measured individuals' total duration unsuccessfully manipulating the box and their solving time, which is the time an individual spent per successful opening after having mastered to open the first lid, therefore representing a measure of a subject's learning efficiency for the new motor task.

Figure 1: The food extraction task: Body width of a mouse lemur corresponds to the width of one compartment (5 x 4.5cm).

String pulling task

In this second novel problem-solving task subjects had to pull a banana piece attached to a string (20cm length) within reach, with only the end of the string inside their test cage, the main part behind the cage mesh on a testing table attached to the cage (Figure 2). Subjects were tested for a maximum of 20 minutes in which we recorded whether a subject succeeded to pull the reward within reach (success y/n) and their success latency from the first moment they payed attention to the string. For subjects that did not succeed we used the total duration of testing, starting with the first attention (i.e. capped latencies).

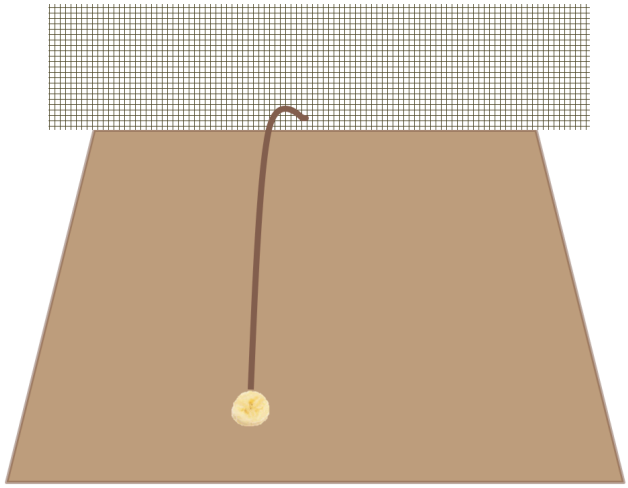


Figure 2: The string-pulling task.

Repeatability and order of testing

All subjects were first tested with the food extraction task and then with the string-pulling task, either in the same night or within the following two nights after the first task. We tested subjects repeatedly in the tasks when recaptured in the given time window (see Table 1).

Table 1: Sample sizes for the repeatability tests

	Food extraction task	String pulling task
10 to 30 days repeat	13	12
1 year repeat	22	23
Contextual consistency	96	

Reference
Huebner, F, Fichtel, C, Kappeler, PM. 2018. Linking cognition with fitness in a wild primate: Fitness correlates of problem-solving performance and spatial learning ability…(follows if accepted in same phil trans issue)

Acknowledgements
We are very thankful to Bruno Tsiverimana, Léonard Razamanantsoa and the rest of the Kirindy research station team for their support in the field. Also, we thank Lynne Werner for helping with the video analyses.

19. SI Klein et al. 2017

Author's name and affiliation:
Simon Klein[1,2], Cristian Pasquaretta[1], Andrew Barron[2], Jean-Marc Devaud[1], Mathieu Lihoreau[1]

[1]Research Center on Animal Cognition (CRCA), Center of Integrative Biology (CBI), University of Toulouse; CNRS, UPS, France
[2]Department of Biological Sciences, Macquarie University, Sydney, NSW,

Author's contribution:
S.K. and M.L. conceived the study and designed the methodology; S.K. collected the data; S.K. and C.P. analysed the data; S.K., C.P., A.B.B., J.M.D. and M.L. wrote the manuscript

Phil trans authors requested:
Simon Klein

Data sharing (full dataset or summary data):
Full datasets (datasets: Klein_bee_spatialConfig ; Klein_bee_foragingBout) will be shared.

Methods. We collected data on spatial learning in bumblebees (*Bombus terrestris*). The task was to find the shortest route to visit four flowers once before returning to the nest over 20 consecutive trials on three different arrays of flowers presented successively to the same bee.

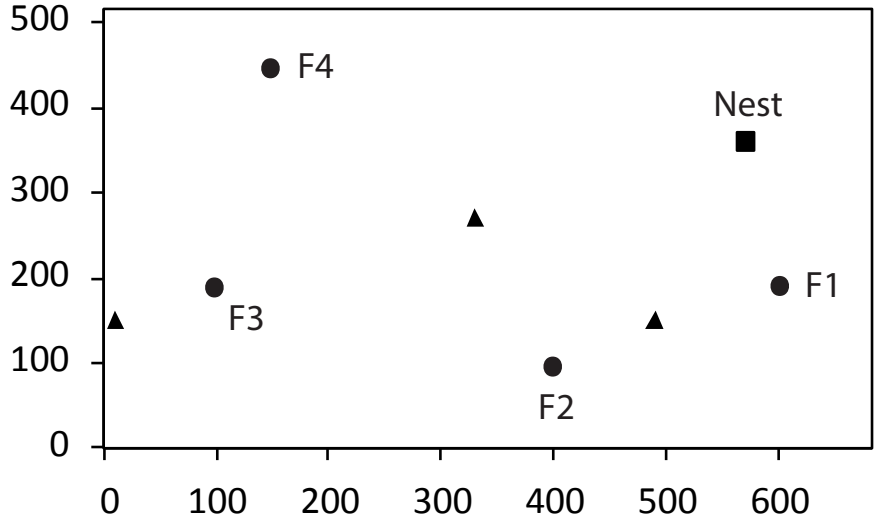
Experimental setup. Experiments were carried out in an indoor flight room with white walls (length: 683 cm, width: 516 cm, height: 250 cm) with controlled illumination, in Spring and Autumn 2015. We used two bumblebee colonies. Each colony was stored in a nest box with a transparent tube and shutters at the entrance to control bee traffic. Bees were individually marked with numbered plastic tags within a day of emergence from pupae in order to monitor their complete foraging history.

Cognitive test. We tested 29 bees, each bee on a different day.

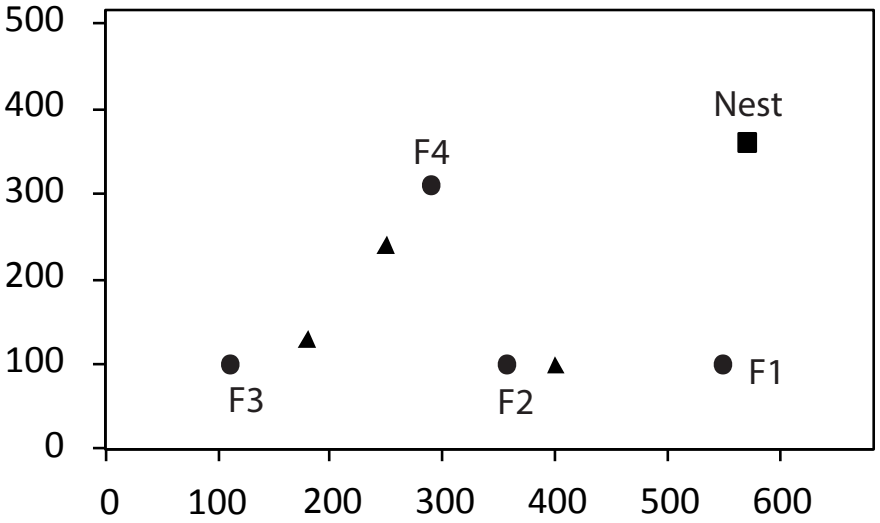
Bees were individually pre-trained to collect sucrose solution from the four artificial flowers arranged in a linear array, located in the middle of the room. Flower rewards were refilled ad libitum with 10 mL of sucrose solution. The mean volume of sucrose solution ingested by a given bee during three successive trials was used to estimate its crop capacity.

Bees were then observed foraging for 20 consecutive trials in the first array of flowers (figure 1A), in the second array of flowers (figure 1B), and in the third array of flowers (figure 1C). During each trial, each flower contained a sucrose reward equivalent to one-fourth of the tested bee's crop capacity and was refilled after each foraging bout. Each array of flowers was also characterised by a unique combination of 3D visual landmarks (figure 1). All departure and arrival times at each flower and at the nest-box were recorded by the experimenter.

A. Flower array 1



B. Flower array 2



C. Flower array 3

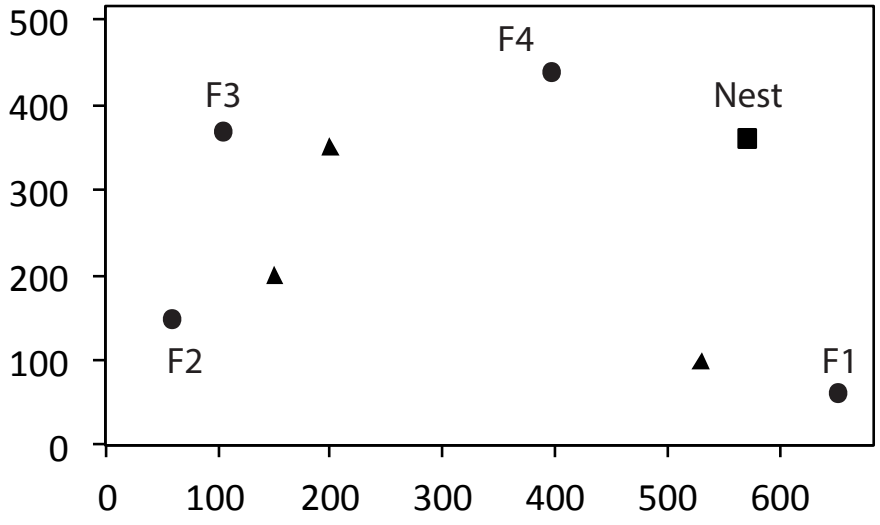


Figure 1. Location of the nest-box (Nest), flowers (F1-F4) and 3D landmarks (black triangles) in the experimental room. Units are in centimetres. See details in Klein et al. (2017).

Cognitive performance:

Dataset : Klein_bee_foragingBout. For each trial, we computed an index OP of route optimisation performance:

IP= 1-|(no/do)-(nmax/dmin)|

where *no* is the number of different flowers visited by the bee, *do* is the estimated total distance travelled by the bee (assuming the bee made straight segments between flowers, and between flowers and the nest), *nmax* is the total number of flowers available in the array, *dmin* is the minimum distance to visit all the flowers once and return to the nest. When IP is 1, the bee uses the optimal route (minimising travel distance between all flowers). If IP is lower than 1, the bee uses a suboptimal route.

Dataset: Klein_bee_spatialConfig. For each individual, we also computed an average OP (Pmean) over the 20 foraging bouts per spatial arrangement of flowers.

Reference

Klein S, Pasquaretta C, Barron AB, Devaud JM, Lihoreau M (2017). Inter-individual variability in the foraging behaviour of traplining bumblebees. Scientific Reports. 7:4561

20. SI Langley & Whiteside (unpublished data)

Authors name and affiliation
Ellis J.G. Langley, Mark A. Whiteside
Centre for Research in Animal Behaviour, Psychology, University of Exeter, Exeter, EX4 4QG, UK

Author's contribution
EJGL and MAW planned the study. EJGL and MAW collected the data. EJGL analysed the data. EJGL and MAW wrote the SI methods.

Phil trans authors requested
EJGL and MAW

Data sharing (full dataset or summary data):
Summary data will be shared.

Methods
General
The study was conducted from March – May 2017 at North Wyke Rothamsted Research Farm, Devon (50□77'N, 3□9'W). We collected data on spatial discrimination task performances of wild-caught adult pheasants (*Phasianus colchicus)*, on two similar, yet distinct tasks; *top-bottom discrimination* and *left-right discrimination* (see *Cognitive testing*). Individuals completed the first task while housed in one of two social treatments (*see Social treatments*) and completed the second task in the alternate social treatment (see Table 1). The type of task and social treatment that individuals received first was determined randomly.

Table 1: Number of participants in each social treatment (*Single* or *Double*) and each spatial discrimination task (*Top-bottom* or *Left-right*)

	Single		***Double***	
	Top-bottom	***Left-right***	***Top-bottom***	***Left-right***
Task 1	9	10	10	10
Task 2	10	10	10	9

The same individuals are represented by grey shading.

Subjects and housing
We captured 58 pheasants (45 females and 13 males) from the wild using baited funnel traps. Eight of these individuals were reared in captivity from the day of hatching for 10 weeks, during which time they were subject to a battery of cognitive tasks (van Horik, Langley, Whiteside, & Madden, 2016) before being released into the wild. The other 46 individuals we caught were birds of unknown rearing history. All birds were ≥10 months old as indicated by their body size and the time of year. Individuals were identifiable by numbered patagial wing tags. Individuals were placed into one of four replicated housing pens, hereafter '*single pens*' (11.4 m x 19 m), ensuring a similar density (approximately 15 individuals) and a similar sex ratio (female : male; 80:20) across pens. Each pen was in visual but not auditory isolation from each other. Each single pen consisted of a large living area that could be separated, during periods of training and testing, into a holding chamber (7.6 m x 3.8 m) and post-testing

chamber (11.4 m x 15.2 m). A separate testing arena (1.27 m x 3.8 m) was attached to the holding area (by a sliding door) and to the post-testing chamber (by a guillotine door). Areas except the testing arena contained refuges, perches and food and water *ad libertum*. The testing arena contained only the testing apparatus.

Social treatments
In the *Single* treatment, single pens remained unchanged. In the *Double* social treatment, we removed panelling that adjoined two of the single pens to allow access between them. Therefore, the social treatments differed in the number of conspecifics that an individual could interact with and the overall area that individuals had access to. On task 1, we determined which two pens remained as the Single treatment and which two pens received the Double treatment, at random. On task 2, pens were assigned the alternate treatment and birds housed in pens that previously received the Double social treatment were returned to their original housing pen. Birds were given three days to habituate to the change in social treatment before testing began on the second task.

Cognitive testing
The test apparatus was a rectangular box (38cm x 14cm x 4cm) presented to birds individually in a testing arena, via a concealed observer. Situated on the 'lid' of the apparatus were two circular wells (diameter 2.8cm), 1.2cm apart; arranged vertically (*Top-bottom discrimination*), or arranged horizontally (L*eft-right discrimination*). A layer of opaque crepe paper covered the wells. For each task, one well contained a mealworm food reward (*correct*), the other well was blocked by a layer of card which could not be pecked through (*incorrect*). Individuals were allowed one choice per trial. During a trial, if an individual made a correct choice, indicated by pecking at the crepe paper of the correct well, they were allowed to consume the food reward before the apparatus was removed. If an individual made an incorrect choice, indicated by pecking at the crepe paper of the incorrect well, the apparatus was immediately removed. Testing on each task lasted five days. On day 1 of each task we assessed their pre-existing bias for a particular well location on 20 trials. If a bird preferentially chose one well (>60% of choices in 20 trials), they were presented with a task in which the alternate well to their preferred well was rewarded on the subsequent test days (2-5). If individuals performed at chance (45-55% choices for either well) then we randomly assigned them a reward location that remained consistent for the subsequent test days (2-5). Testing on days 2-5 consisted of one test session per day, comprising of 10 trials to give 40 trials by the end of each task. Cognitive performances were calculated as the accuracy over 40 trials (number of correct responses / total number responses).

Training
Training individuals to the testing procedures took place while birds were housed in their original pens and lasted five weeks. During the first two weeks, we allowed birds' access to the testing arena by leaving both the entry and exit doors to the testing arena open and continuously baiting the testing apparatus. This allowed individuals to 'discover' and forage within the testing arena in groups. During the third week we trained individuals to the walk-in procedure. All birds were ushered into the holding chamber and were trained to enter the testing arena via the sliding entrance door. This was achieved by initially allowing groups to enter and reducing group sizes until birds were comfortable to enter the testing arena individually. During this process the test apparatus was heavily baited without crepe-paper. Birds were trained to exit the arena into the post-testing area, when the exit door was raised. For the remaining two weeks we trained individuals to peck through crepe paper to obtain a mealworm reward. All wells were rewarded. Participating individuals received the same

number of exposures to the testing apparatus and we ensured that these individuals ‘opened’ the same number of rewarded wells. Some individuals did not habituate to the testing procedures and exhibited behavioural signs of stress if allowed entry to the testing chamber. If at any of the stages individuals failed to engage in the procedures on three consecutive occasions, we ceased in our attempts to train them. These individuals remained in the holding area with conspecifics until testing finished.

21. SI Lihoreau et al. 2012a

Author's name and affiliation:
Mathieu Lihoreau[1,2], Nigel Raine[2,3], Andy Reynolds[4], Ralph Stelzer[2], Ka Lim[4], Alan Smith[4], Juliet Osborne[4], Lars Chittka[2]

[1]Research Center on Animal Cognition (CRCA), Center of Integrative Biology (CBI), University of Toulouse; CNRS, UPS, France
[2]Biological and Experimental Psychology Group, School of Biological and Chemical Sciences, Queen Mary University of London, London, United Kingdom
[3]School of Biological Sciences, Royal Holloway, University of London, Egham, TW 20 OEX, UK
[4]Rothamsted Research, Harpenden, Hertfordshire, United Kingdom

Author's contribution:
ML, NR and LC conceived and designed the experiments. ML, NR, RS and AS performed the experiments. ML, RS and AR analysed the data. AR, KL, AS and JO contributed reagents/materials/analysis tools. ML, NR, AR, AS, JO and LC wrote the paper.

Phil trans authors requested:
Mathieu Lihoreau

Data sharing (full dataset or summary data):
Full datasets (dataset : Lihoreau_bee_plosb) will be shared.

Methods. We collected data on spatial learning in bumblebees (*Bombus terrestris*). The task was to find the shortest route to visit five flowers once and return to the nest over 22 to 37 consecutive trials.

Experimental setup. Experiments were carried out in October 2010, in a flat, open area of mown pasture (approximately 700 x 300 m) on the Rothamsted estate (Hertfordshire, UK). We used one bumblebee colony with a transparent tube and shutters at the entrance to control bee traffic. Bees were individually marked with numbered plastic tags within a day of emergence from pupae in order to monitor their complete foraging history.

Cognitive test:
We tested 7 bees, each on a different day.

Bees were individually pre-trained to collect sucrose solution from the five artificial flowers arranged in a linear array (150 cm length), located 50 m north-west of the nest entrance. Flower rewards were refilled ad libitum with 10 mL of sucrose solution. The mean volume of sucrose solution ingested by a given bee during three successive foraging bouts was used to estimate its crop capacity.

Bees were then observed foraging on the five flowers arranged in a regular pentagon (50 m side, figure 1), until they visited all flowers during at least five consecutive trials (total of 22-37 trials, N = 7 bees). Each flower contained a sucrose reward equivalent to one-fifth of the test bee's crop capacity and was refilled after each trial. All departure and arrival times at the nest-box were recorded by an experimenter. Flower visits were automatically recorded using

motion-activated webcams on flowers.

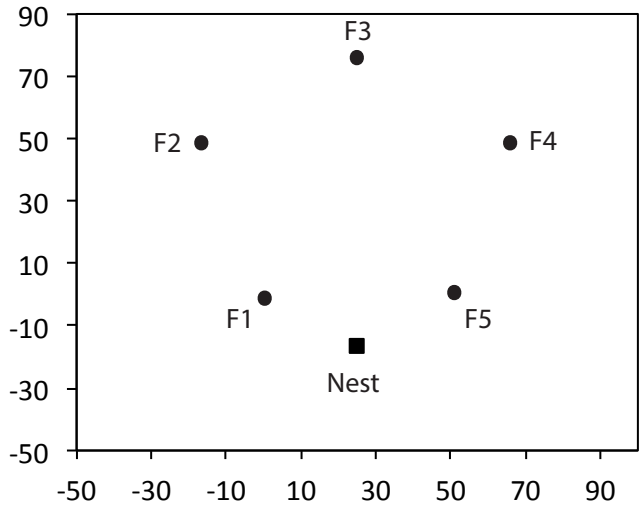


Figure 1. Location of the nest-box (Nest) and flowers (F1-F5) in the field. Units are in meters. See details in Lihoreau et al. (2012a).

Cognitive performance:

Time-coded video clips from all flowers allowed us to reconstruct the visitation sequence for every foraging bout of each bee. For each trial, we computed an index OP of route optimisation performance:

OP= 1-|(no/do)-(nmax/dmin)|

where *no* is the number of different flowers visited by the bee, *do* is the estimated total distance travelled by the bee (assuming the bee made straight segments between flowers, and between flowers and the nest), *nmax* is the total number of flowers available in the array, *dmin* is the minimum distance to visit all the flowers once and return to the nest. When OP is 1, the bee uses the optimal route (minimising travel distance between all flowers). If OP is lower than 1, the bee uses a suboptimal route.

Reference

Lihoreau M, Raine NE, Reynolds AM, Stelzer RJ, Lim KS, Smith AD, Osborne JL, Chittka L (2012) Radar tracking and motion sensitive cameras on flowers reveal the development of pollinator multi-destination routes over large spatial scales. PLoS Biology 10:e1001392.
SI Lihoreau et al. 2012b (dataset : Lihoreau_bee_bioLett) :

22. SI Lihoreau et al. 2012b

Author's name and affiliation:
Mathieu Lihoreau[1,2], Lars Chittka[2], Le Comber Steven[2], Nigel Raine[2,3]

[1]Research Center on Animal Cognition (CRCA), Center of Integrative Biology (CBI), University of Toulouse; CNRS, UPS, France
[2]Biological and Experimental Psychology Group, School of Biological and Chemical Sciences, Queen Mary University of London, London, United Kingdom
[3]School of Biological Sciences, Royal Holloway, University of London, Egham, TW 20 OEX, UK

Author's contribution:
ML, LC and NR designed the experiments. ML performed the experiments and analysed the data. All authors wrote the paper.

Phil trans authors requested:
Mathieu Lihoreau

Data sharing (full dataset or summary data):
Full datasets (dataset : Lihoreau_bee_bioLett) will be shared.

Methods. We collected data on spatial learning in bumblebees (*Bombus terrestris*). The task was to find the shortest route to visit six flowers once before returning to the nest over 80 consecutive trials.

Experimental setup. Experiments were carried out in March 2010, in an indoor flight room (approximately 8.7 x 7.3 x 2 m) with controlled illumination. We used one bumblebee colony with a transparent tube and shutters at the entrance to control bee traffic. Bees were individually marked with numbered plastic tags within a day of emergence from pupae in order to monitor their complete foraging history.

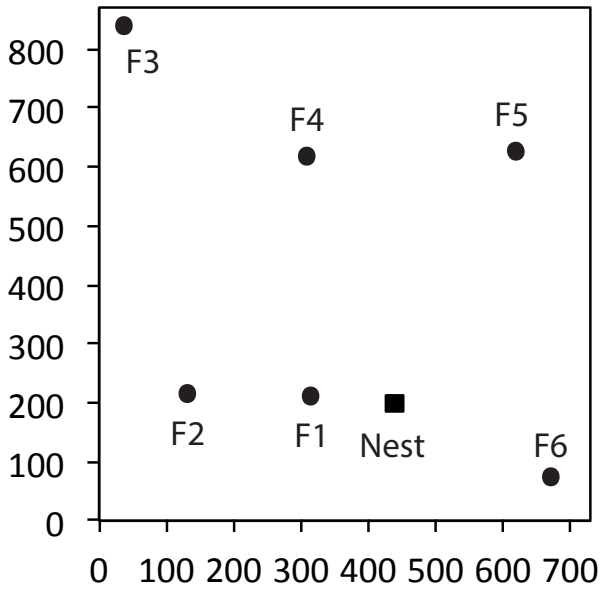


Figure 1. Location of the nest-box (Nest) and flowers (F1-F6) in the experimental room. Units

are in centimetres. See details in Lihoreau et al. (2012b).

Cognitive test:
We tested 8 bees, each on a different day.

Bees were individually pre-trained to collect sucrose solution from the six artificial flowers arranged in a linear array, located 1 m of the nest entrance. Flower rewards were refilled ad libitum with 10 mL of sucrose solution. The mean volume of sucrose solution ingested by a given bee during three successive foraging bouts was used to estimate its crop capacity.

The bees were then observed foraging on the six flowers, for 80 consecutive trials. The flowers were distributed in order to maximise the discrepancy between the shortest possible route and the route linking all nearest unvisited flowers (figure 1). Each flower contained a sucrose reward equivalent to one-sixth of the test bee's crop capacity and was refilled after each foraging bout. All departure and arrival times at each flower and at the nest-box were recorded by an experimenter.

Cognitive performance:

For each trial, we computed an index OP of route optimisation performance:

$$OP = 1-|(no/do)-(nmax/dmin)|$$

where *no* is the number of different flowers visited by the bee, *do* is the estimated total distance travelled by the bee (assuming the bee made straight segments between flowers, and between flowers and the nest), *nmax* is the total number of flowers available in the array, *dmin* is the minimum distance to visit all the flowers once and return to the nest. When OP is 1, the bee uses the optimal route (minimising travel distance between all flowers). If OP is lower than 1, the bee uses a suboptimal route.

Reference
Lihoreau M, Chittka L, Le Comber SC, Raine NE (2012) Bees do not use nearest-neighbour rules for optimization of multi-location routes. Biology Letters 8:13-16.

23. SI Lihoreau et al. 2011

Author's name and affiliation:
Mathieu Lihoreau[1,2], Lars Chittka[2], Nigel Raine[2,3]

[1]Research Center on Animal Cognition (CRCA), Center of Integrative Biology (CBI), University of Toulouse; CNRS, UPS, France
[2]Biological and Experimental Psychology Group, School of Biological and Chemical Sciences, Queen Mary University of London, London, United Kingdom
[3]School of Biological Sciences, Royal Holloway, University of London, Egham, TW 20 OEX, UK

Author's contribution:
ML, LC and NR designed the experiments. ML performed the experiments and analysed the data. All authors wrote the paper.

Phil trans authors requested:
Mathieu Lihoreau

Data sharing (full dataset or summary data):
Full datasets (dataset: Lihoreau_bee_functEcol) will be shared.

Methods. We collected data on spatial learning in bumblebees (*Bombus terrestris*). The task was to find the shortest route to visit five flowers once before returning to the nest over 40 consecutive trials.

Experimental setup. Experiments were carried out in May 2010, in an indoor flight room (approximately 8.7 x 7.3 x 2 m) with controlled illumination. We used one bumblebee colony with a transparent tube and shutters at the entrance to control bee traffic. Bees were individually marked with numbered plastic tags within a day of emergence from pupae in order to monitor their complete foraging history.

Cognitive test:
We tested 10 bees, each on a different day.

Bees were individually pre-trained to collect sucrose solution from the six artificial flowers arranged in a linear array, located 1 m of the nest entrance. Flower rewards were refilled ad libitum with 10 mL of sucrose solution. The mean volume of sucrose solution ingested by a given bee during three successive trials was used to estimate its crop capacity.

The bees were then observed foraging on the five flowers, for 40 consecutive trials. Each flower contained a sucrose reward equivalent to one-fifth of the test bee's crop capacity and was refilled after each trial. All departure and arrival times at each flower and the nest-box were recorded by an experimenter.

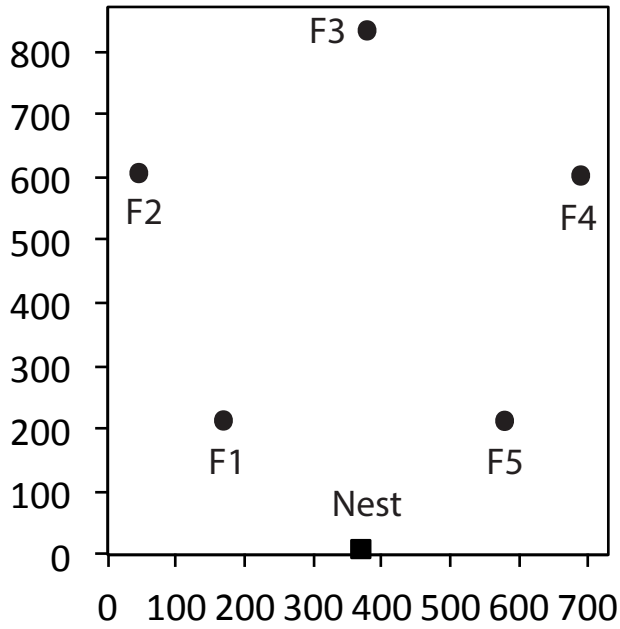


Figure 1. Location of the nest-box (Nest) and flowers (F1-F5) in the experimental room. Units are in centimetres. See details in Lihoreau et al. (2011).

Cognitive performance:

For each trial, we computed an index OP of route optimisation performance:

OP= 1-|(no/do)-(nmax/dmin)|

where *no* is the number of different flowers visited by the bee, *do* is the estimated total distance travelled by the bee (assuming the bee made straight segments between flowers, and between flowers and the nest), *nmax* is the total number of flowers available in the array, *dmin* is the minimum distance to visit all the flowers once and return to the nest. When OP is 1, the bee uses the optimal route (minimising travel distance between all flowers). If OP is lower than 1, the bee uses a suboptimal route.

Reference

Lihoreau M, Chittka L, Raine NE (2011) Trade-off between travel distance and prioritization of high reward sites in traplining bumblebees. Functional Ecology 25:1284-1292.

24. SI Matzel et al. Attention dataset

We collected data on individual differences in attention using 26 male CD-1 mice. The mice started the tests at approximately 90 days of age in the Spring of 2012, and the testing lasted for 3 months (in Piscataway, USA). Mice were group housed among siblings prior to weaning at 21 days of age, and singly housed thereafter. They were singly housed in clear shoebox cages inside a temperature-controlled colony room with a 12:12 h light–dark cycle. Unless noted, the animals had free access to food and water.

We used four attention tests: Mouse Stroop Test, T-Maze Reversal, Coupled Latent Inhibition, and Dual Radial Arm Maze (administered in this order). The tests were designed to differently capture two types of attention: Attention against external sources of interference (interference from environmental cues) and attention against internal sources of interference (interference from memory and habit). The Mouse Stroop Test and the Dual Radial Arm Maze were mainly external, and the Coupled Latent Inhibition and the T-Maze were mainly internal.

Before the start of the study, we handled all mice for 14 days prior to the start of any testing to minimize noise in the data from stress. (Handling consisted of holding a mouse on the palm of an experimenter's hand, and systematically walking it around the laboratory.) We also had a session of adaptation in the apparatus on the day prior to the start of each test. For the Mouse Stroop Test, T-Maze Reversal, and Dual Radial Arm Maze, we food deprived the mice by giving them only 90 min of access to food daily, beginning on the day prior to training. The data in the Mouse Stroop Test, Coupled Latent Inhibition, and Dual Radial Arm Maze were multiplied by -1, so higher values always mean better attention for all tests. Here we summarize the procedures for each test. They are also described in more detail in Sauce et al. (2014), together with results of mice's performance.

Mouse Stroop Test
Mice were first required to associate meaning to odor and visual cues, analogous to the way the human Stroop Test requires subjects to know how to identify words and colors. For this, we trained the mice in two tasks: odor discrimination and visual discrimination. In the odor discrimination task, mice had to use a specific odor cue to find food. The task was administered in a square box of black Plexiglas, where three of the box's four corners always contained cups, and the fourth corner served as a start location. In this task, the animal's goal was to locate the accessible food using the odor of mint as the discriminative stimulus. In the visual discrimination task, mice had to use a specific visual cue (pattern of illuminated holes forming an X) to find the food, but in a different box. Everything else was the same as for the odor discrimination. After training separately in the odor and visual discrimination tasks, we then conducted the Mouse Stroop Test. It consisted of a complex discrimination task that requires mice to ignore interference from one of the learned target cues and maintain attention on the context-appropriate cue. The context (the two different training boxes) determined if the relevant cue (the one that marks the available food) is either visual (X) or odor (MINT). For this purpose, we conducted the tests in the visual box with both the previously trained visual cues (with X as the target), and odor cues (with MINT as the task-relevant external distractor). We used the average number of errors in the visual box of the Mouse Stroop Test as a measure of attention.

T-Maze Reversal
We first trained the mice in a reinforced alternation task where they must alternate their foraging (for a food reward) between two arms. Then, reversal training began, wherein food

was always located in the same arm. This reversal training required animals to ignore the previously learned response and maintain attention to the new task requirements. In other words, the animal must resist a source of interference that originates internally (i.e., the prior learning). The apparatus consisted of a start arm that intersected at its extremity with two choice arms, forming a ''T'' shape. To help the mice distinguish between arms, one of the arms' walls had vertical white stripes, and the other had horizontal white stripes. If an incorrect choice was made, we allowed the animal to correct its mistake and find the food in the other arm. After the correct choice was made, we placed the animal back in the start area and waited 20 s for the following trial. To specifically assess animals' ability to reverse (i.e., ignore a previously learned response rule, or resist this source of internal interference), we divided the results from the reversal by the average number of correct choices during the same period (trials 5–12, out of 24) of the reinforced alternation task. Hence, this measure reflected an animal's reversal performance relative to its respective normal performance (and thus should be relatively independent of learning ability per se). We used this as a measure of mice's attention (better attention = higher values).

Coupled Latent Inhibition

In latent inhibition, an animal is repetitively exposed to a stimulus (that will later serve as a CS) that has no explicit meaning (i.e., it is presented alone). Subsequently, it is difficult for that animal to learn to associate that stimulus with a second stimulus (i.e., associative learning is impaired). Here, we used a Coupled Latent Inhibition procedure that, in principle, could assess variation in attention independently of variation in learning. First, we conducted a fear conditioning task (tone-shock pairings) to determine each mouse's learning rate in the absence of interference (i.e., no prior exposure to the CS). Then, we conducted a fear conditioning task to determine their rate of learning of a light-shock association after extensive latent inhibition trials with non-reinforced exposure to the light (i.e., interference from prior experience). In this later case, the animals had to overcome the habit of ignoring the light, and maintain attention to the new relevance of this stimulus (i.e., its relationship to the shock). Here, the measure of the effects of latent inhibition (interference) was reflected as the difference between the rate of learning of the tone-shock and lightshock associations. For both Fear Conditioning and Latent Inhibition, we calculated the time of CS freezing by measuring the time spent freezing during the 20 s of tone, and subtracted the time spent freezing during 20 s before the tone (the latter a measure of context freezing). We then subtracted, for each animal, the average value during acquisition of fear to CS1 during Fear Conditioning (trials 2–4) from the average value during the same period (trials 2–4) of acquisition of fear to CS2 during Latent Inhibition. The result was the value for Coupled Latent Inhibition, which is a measure of attention. A good performance in attention requires values after latent inhibition to be close to (or higher than) the performance in fear conditioning before latent inhibition.

Dual Radial Arm Maze

We assessed animals' ability to operate simultaneously on two related sets of guidance cues. We first trained the mice in two different (visually distinct) eight-arm radial mazes located in a single testing room (which thus shared extra-maze visual cues). In order to efficiently find food, each maze requires an animal to use spatial cues (distributed around the maze) to guide its search and/or to maintain a memory of arms that have been visited within a trial. After reaching asymptotic (near errorless) performance in each maze, the attentional phase of testing began. During this phase, animals had to alternate choices between the two mazes. The two mazes were located in a single room, so the spatial cues were common to both mazes. Since animals must maintain a memory of the cues segregated according to the appropriate reference maze, the test required animals to maintain attention to the spatial cues relevant to the maze in

which it was currently in, and ignore external interference from the cues appropriate for the other maze. We recorded the number of errors that an animal made, and used the average of the three trials (each a sum of errors in the black and grey mazes) during the Dual Radial Arm Maze test as a measure of their attention.

Reference

Sauce, B., Wass, C., Smith, A., Kwan, S., & Matzel, L. D. (2014). The external–internal loop of interference: Two types of attention and their influence on the learning abilities of mice. *Neurobiology of Learning and Memory*, *116*, 181–192. http://doi.org/10.1016/j.nlm.2014.10.005

25. SI Matzel et al. Learning dataset

We collected data on individual differences in learning ability using 56 male CD-1 mice. The mice started the tests (in several independent replications) at approximately 90 days of age in the fall of 2001, and each replication required about 30 days to complete (performed in Piscataway, USA). Mice arrived in our laboratory at approximately 70 days of age and were singly housed in clear shoebox cages inside a temperature-controlled colony room with a 12:12 h light–dark cycle. Unless noted, the animals had free access to food and water. We handled all mice for 14 days prior to the start of any testing to minimize noise in the data from stress. (Handling consisted of holding a mouse on the palm of an experimenter's hand, and systematically walking it around the laboratory.)

Using previously reported methods (Kolata et al., 2005; Matzel et al., 2003; Matzel et al., 2006), all animals were tested in a series of five independent learning tasks: Lashley III maze (egocentric maze learning), passive avoidance (where stepping off a safe platform initiated an aversive compound of light and noise), spatial water maze (where the animals used spatial cues to find a hidden escape platform in pool of water), odor discrimination (where a distinct odor guided the animals search for food), and cued fear conditioning (where animals learned to associate a tone with footshock). By design, these tasks place unique sensory, motor, motivational and information processing demands on the animals.

These five tests were administered in an order that separated tasks that have similar patterns of action (e.g. activity or passivity). This ordering prevented excessive physical strain and minimizes any potential cross-task influences due to motivational factors. The order in which animals were tested was: Lashley maze, passive avoidance, odor discrimination, spatial water maze, and fear conditioning. A different experimenter tested the animals on each of the learning tasks.

In all learning tasks, the animals' performance was assessed during the acquisition phase of learning (i.e., prior to reaching their stable, asymptotic level of performance). Thus the dependent measure for each task was analogous to the animals' rate of learning on that task, and these measures of each individual's performance could be ranked (through the application of exploratory factor analysis and the resultant factor scores; see below) relative to other animals in the sample. To quantify an animal's performance in tasks in which there were multiple training/test trials, performance during trials that fell within the acquisition phase were averaged. In tasks in which there was only one test trial (i.e. fear conditioning and passive avoidance), training parameters were used that were previously determined to result in sub-asymptotic responding by most animals (Matzel et al., 2003), and as such, performance on this single test trial were in part a reflection of differences in animal's rate of learning.

[what follows are detailed methods; since these methods have been published many times previously, they can be omitted if you need to be more brief]

Spatial Water Maze. This task requires animals to locate a submerged platform in a round pool of opaque water. A round black pool (140 cm diameter, 56 cm deep) was filled to within 24 cm of the top with water made opaque by the addition of a nontoxic, water soluble black paint. A hidden 11 cm diameter perforated black platform was in a fixed location 1.5 cm below the surface of the water midway between the center and perimeter of the pool. The pool was enclosed in a ceiling-high black curtain on which five different shapes (landmark cues) were variously positioned at heights (relative to water surface) ranging from 24-150 cm. Four of

these shapes were constructed of strings of white LEDs (spaced at 2.5 cm intervals) and include an "X" (66 cm arms crossing at angles 40o from the pool surface), a vertical "spiral" (80 cm long, 7 cm diameter, 11 cm revolutions), a vertical line (31 cm) and a horizontal line (31 cm). The fifth cue was constructed of two adjacent 7-watt light bulbs (each 4 cm diameter). A video camera was mounted 180 cm above the center of the water surface. These cues provided the only illumination of the maze, totaling 172 lux at the water surface.

On the day prior to training, each animal was confined to the escape platform for 5 min. Training was conducted on the two subsequent days. On Day 1 of training, animals were started from one of three unique locations on each of five trials. The pool was conceptually divided into four quadrants, and one starting point was located in each of the three quadrants that did not contain the escape platform. The starting point on each trial alternated between the three available quadrants. An animal was judged to have escaped from the water (i.e., located the platform) at the moment at which all four paws were situated on the platform, provided that the animal remained on the platform for at least 5 sec. Each animal was left on the platform for a total of 20 sec, after which the trial was terminated. Six trials were spaced at 10 min intervals, during which time the animals were held in their home cages. On each trial, a 90 sec limit on swimming was imposed, at which time any animal that had not located the escape platform was placed onto the platform by the experimenter, where it remained for 20 sec. The time it took for the animal to escape (latency) as well as the distance traveled (path length) to reach the platform were recorded.

Lashley III Maze. The Lashley III maze consisted of a start box, four interconnected alleys and a goal box containing a food reward. Previous studies have shown that over successive trials, the latency of rats to locate the goal box decreased, as does their number of errors (i.e., wrong turns or retracing). A Lashley III maze scaled for mice was constructed of black Plexiglas and a goal box marked by white electrical tape was located in the rear portion of the maze where 45 milligram BioServe (rodent grain) pellet served as a reinforcer. Illumination was 80 lux at the floor of the maze. The maze was isolated behind a shield of white Plexiglas to prevent the use of extra-maze landmark cues.

For the two days prior to training, the mice' access to food was limited to 60 min per day at the end of the light cycle. The food-deprived mice were acclimated and trained on two successive days. On the day prior to acclimation, all animals were provided with three food pellets (the reinforce) in their home cages to familiarize them with the novel reinforcer. On the acclimation day, each mouse was placed in the four alleys of the maze, but the openings between the alleys were blocked so that the animals could not navigate the maze. Each animal was confined to the start box and subsequent two alleys for 4 min, and for 6 min in the last (goal) alley, where three food pellets were present in the goal box. This acclimation period promotes stable and high levels of activity on the subsequent training day. On the training day, each animal was placed in the start box and allowed to traverse the maze until it reached the goal box and consumed the single food pellet present in the cup (a 1 cm depression in the floor at the rear of the box). Upon consuming the food, the animal was returned to its home cage for a 20 min interval (ITI) during which the apparatus was cleaned. After the ITI, the mouse was returned to the start box to begin the next trial, and this sequence was repeated for five trials. The latency and errors (i.e., a turn in an incorrect direction, including those which result in path retracing) to enter the goal box were recorded on each trial.

Associative Fear Conditioning. In this task, mice received a tone (CS) paired with a mild foot shock (US). The training chamber (16.5 x 26.5 x 20 cm) was brightly illuminated (100 lx), had

clear Plexiglas walls, and parallel stainless-steel rods (5mm, 10mm spacing) forming the floor. The auditory stimulus (60dB, 2.9 kHz) was delivered by a piezoelectric buzzer.

On Day 1 subjects were acclimated to the training context for a 20 min. On Day 2 subjects received an 18 min training session in the training chamber. All training sessions were videotaped for subsequent offline scoring. Subjects received three tone/ shock presentations at 4, 10 and 16 min into the session. The CS presentation consisted of a pulsed (.7 sec on .3 sec off) 20 sec tone. Immediately following the tone offset, the shock US (0.6-mA, constant-current foot shock) was presented for 500 msec. Freezing was measured during the 20 sec before (baseline freezing), during (tone freezing) and after (post shock freezing) the 20 sec tone presentation. A measure for freezing during the training period was calculated by subtracting the time spent freezing in baseline from the time spent freezing during the tone.

Odor Discrimination. Rodents rapidly learn to use odors to guide appetitively-reinforced behaviors. In a procedure based on one designed for rats (Sara, Roullet, & Przybyslawski, 2001), mice learned to navigate a square field in which unique odor-marked (e.g., almond, lemon, mint) food cups were located in three corners. Although food was present in each cup, it was accessible to the animals in only one cup (marked by mint). An animal was placed in the empty corner of the field, after which it explored the field and eventually retrieved the single piece of available food. On subsequent trials, the location of the food cups was changed, but the accessible food was consistently marked by the same odor (mint). On successive trials, animals required less time to retrieve the food and made fewer approaches (i.e., "errors") to those food cups in which food was not available. Using this procedure, errorless performance was typically observed within three to four training trials.

A black Plexiglas 60 cm square field with 30 cm high walls was located in a dimly lit (10 fc) testing room with a high ventilation rate (3 min volume exchange). Three 4 x 4 x 2.0 cm (l, w, h) aluminum food cups were placed in three corners of the field. A food reinforcer (30 mg portions of chocolate flavored puffed rice) was placed in a 1.6 cm deep, 1 cm diameter depression in the center of each cup. The food in two of the cups was covered (1.0 cm below the surface of the cup) with a wire mesh so that it was not accessible to the animal, while in the third cup (the "target" cup), the food could be retrieved and consumed. A cotton-tipped laboratory swab, located between the center and rear corner of each cup, extended vertically 3 cm from the cups' surface.

Immediately prior to each trial, fresh swabs were loaded with 25 µl of lemon, almond, or mint odorants (McCormick flavor extracts). The mint odor was always associated with the target food cup. It should be noted that in pilot studies, the odor associated with food was counterbalanced across animals and no discernible differences in performance could be detected in response to the different odors.

On test day, animals received four training trials in the field with all three food cups present. On each trial, a mouse was placed in the empty corner of the field. On Trial 1, the reinforcing food (one piece of chocolate flavored puffed rice) was available to the animal in the cup marked by mint odor. An additional portion of food was placed on the top surface of the same cup for the first trial only. The trial continued until the animal retrieved and consumed the food from the target cup, after which the animal was left in the chamber for an additional 20 sec and then returned to its home cage to begin a 6 min ITI. On Trials 2-4, the location of the food cups was rearranged, but the baited cup remained consistently marked by the mint odor. Both the

corner location of the mint odor and its position relative to the remaining odors was changed on each trial. On each trial, the latency to retrieve the food and errors was recorded. An error was recorded any time an animal made contact with an incorrect cup, or its nose crossed a plane parallel to the perimeter of an incorrect cup. Similarly, an error was recorded when an animal sampled (as above) the target cup but did not retrieve the available food.

Passive Avoidance. A chamber illuminated by dim (< 20 lux) red light was used for training and testing. Animals were confined to a circular ("safe") chamber (10 cm diameter, 8 cm high). The walls and floor of this chamber were white, and the ceiling was translucent orange. The floor was comprised of plastic rods (2 mm diameter) arranged to form a pattern of 1 cm square grids. A clear exit door (3 cm square) was flush with the floor of the safe compartment, and the door was able to slide horizontally to open or close the compartment. The bottom of the exit door was located 4 cm above the floor of a second circular chamber (20 cm diameter, 12 cm high). This "unsafe" chamber had a clear ceiling and a floor comprised of 4 mm wide aluminum planks that formed a pattern of 1.5 cm square grids oriented at a 45o angle relative to the grids in the safe compartment. When an animal stepped from the safe compartment through the exit door onto the floor of the unsafe compartment, a compound aversive stimulus comprised of a bright (550 Lux) white light and "siren" (60 dB above the 50 dB background) was initiated. Animals learn to suppress movement to avoid contact with aversive stimuli. This "passive avoidance" response is exemplified in step-down avoidance procedures, where commonly, an animal is placed on a platform, whereupon stepping off of the platform it encounters a foot shock. Following just a single encounter with shock, animals are subsequently reluctant to step off of the safe platform. The animals' reluctance to leave the platform is believed to not reflect fear, because typical fear responses are not expressed in animals engaged in the avoidance response. Upon stepping off the platform, animals here were exposed to a compound of bright light and loud oscillating noise rather than shock, so as not to duplicate stimuli between tasks (see fear conditioning, above). Like more common procedures, our variant of this task supports learning after only a single trial (i.e., subsequent step-down latencies will be markedly increased).

Animals were placed on the platform behind the exit blocked by the Plexiglas door. After 4 min of confinement, the door was retracted and the latency of the animal to leave the platform and make contact with the grid floor was recorded. Prior to training, baseline step-down latencies typically ranged from 8-20 sec. Upon contact with the floor, the door to the platform was closed and the aversive stimulus (light and noise) was presented for 4 sec, at which time the platform door was opened to allow animals to return to the platform, where they were again confined for 5 min. At the end of this interval, the door was opened and the latency of the animal to exit the platform and step onto the grid floor (with no aversive stimulation) was recorded. The ratio of post-training to pre-training step-down latencies was calculated for each animal and this served to index learning. We have determined that asymptotic performance is apparent in group averages following 2-3 training trials; thus performance after a single trial reflects, in most instances, sub-asymptotic learning.

Upon completion of the above tests, we determined the degree to which animal's performance on each task was a reliable index of its ability. To address this, eight of the original 56 animals were trained and tested on a second series of learning tasks that were variants of each of the above. Each of the tasks in the second battery required new learning, although the nature of the tasks and the underlying processes were nominally identical to those which comprised the first series of tests. With data obtained from animals tested in each of the two batteries it was possible to assess the degree of consistency of individual animals' ranks on each of two

analogous tasks, as well as the degree to which individuals' aggregate performances (i.e., average ranks) were correlated across the two series of tests.
Upon completion of the initial battery, animals began a second series of tests. Modifications of the tasks were as follows: 1. The black Lashley III maze was replaced with a white maze that required a different route to efficiently retrieve the food reinforcer; 2. For passive avoidance, animals were trained in a distinct context and the safe platform was white (c.f., black). Furthermore, an odor (28 g Vick's VapoRub) was added to the chamber to distinguish it from the chamber that was previously used. 3. In the water maze, the spatial cues were replaced by a new set of geometric shapes located at different coordinates, the escape platform was moved to a different quadrant of the maze, and start locations were changed; 4. For odor discrimination, three new odors (i.e., rum, anise, coconut [target]) were used as discriminative cues, the pattern of start locations were changed, and the training context was black (c.f., white); 5. A new training context was employed for fear conditioning, and a flashing light (250 msec on/250 msec off) located in the top center of each box served as the CS.

References

Kolata, S., Light, K., Townsend, D. A., Hale, G., Grossman, H., & Matzel, L. D. (2005). Variations in working memory capacity predict individual differences in general learning abilities among genetically diverse mice. Neurobio.Learn.Mem., 84, 242-246.

Matzel, L. D., Han, Y. R., Grossman, H., Karnik, M. S., Patel, D., Scott, N., . . . Gandhi, C. C. (2003). Individual differences in the expression of a "general" learning ability in mice. Journal of Neuroscience, 23, 6423-6433.

Matzel, L. D., Townsend, D. A., Grossman, H., Han, Y. R., Hale, G., Zappulla, M., . . . Kolata, S. (2006). Exploration in outbred mice covaries with general learning abilities irrespective of stress reactivity, emotionality, and physical attributes. Neurobio.Learn.Mem., 86, 228-240.

Sara, S. J., Roullet, P., & Przybyslawski, J. (2001). Consolidation of memory for odor-reward association: B-adrenergic receptor involvement in the late phase. Learning and Memory, 6, 88-96.

26. SI Nawroth et al. 2013

Author's name and affiliation:
Nawroth Christian[1], Ebersbach Mirjam[2], von Borell Eberhard[3]

[1]Leibniz Institute for Farm Animal Biology, Institute of Behavioural Physiology, Dummerstorf, Germany
[2]University of Kassel, Institute of Psychology, Kassel, Germany
[3]University of Halle-Wittenberg, Institute of Agricultural and Nutritional Sciences, Halle, Germany

Author's contribution:
CN, ME and EvB planned the study. CN collected the data. CN analysed the data. CN wrote the SI methods.

Phil trans authors requested:
None

Data sharing (full dataset or summary data):
Summary data will be shared.

Methods:
General. We collected data on the use of human-given cues in domestic pigs (*Sus scrofa*) during 4 different experiments carried out in Halle (Germany) over a period of 4 weeks. Each experiment ran over 1 week.

For all experiments, data were collected using an object choice task. For training and testing, the pigs were separated in a compartment (starting area, 250 cm x 400 cm) adjacent next to their home pen. An experimenter was located in an adjacent test area (250 cm x 400 cm). Starting and test area were connected by a corridor leading to an entrance (50 cm). In all experiments, two metal food bowls (diameter: 20 cm, height: 5 cm) were placed 150 cm away from the entrance in the test area and 140 cm (except for Experiment 2) apart from each other while the experimenter was in a kneeling position about 30 cm behind the midline of both bowls. When a test subject entered the test area, it was free to explore the location and to choose one of the two bowls. Habituation of subjects to the test area and detailed training procedures are described in Nawroth et al. (2014).

Experiment 1
Seventeen domestic pigs (8 males, 9 females, 7 weeks) participated in this experiment. The experimenter baited one of the two bowls surreptitiously with a piece of food. After baiting, the test subject was allowed to enter the test area from the starting area. Subjects received one of four different test conditions:

a) proximal dynamic-sustained pointing and gaze (PDS-G) - The experimenter kneeled between the two bowls and as soon as the subject entered the corridor, he pointed and turned his head towards the baited bowl until the subject made a choice. The distance between the tip of the index finger and the baited bowl was about 30 cm.

b) proximal momentary pointing (PM) - The experimenter kneeled between the two bowls and as soon as the subject entered the corridor, he pointed towards the baited bowl for about one

second or as long as the subject was still in the corridor. Pigs never entered the test area while the gesture was still being administered. The distance between the tip of the index finger and the baited bowl was about 30 cm.

c) distal dynamic-sustained pointing (DDS) - The experimenter stood between the two bowls and as soon as the subject entered the corridor, he pointed towards the baited bowl until the subject made a choice. The distance between the tip of the index finger and the baited bowl was about 80 cm.

d) distal momentary pointing (DM) - The experimenter stood between the two bowls and as soon as the subject entered the corridor, he pointed towards the baited bowl for about one second or as long as the subject was still in the corridor. Pigs never entered the test area while the gesture was still being administered. The distance between the tip of the index finger and the baited bowl was about 80 cm.

Subjects received five test sessions of 16 trials each (four trials for every condition in each session) with a total of 20 trials of each condition. Detailed methods and data are published in Nawroth et al. (2014).

Experiment 2

Fifteen domestic pigs (7 males, 8 females, 8 weeks) participated in this experiment. The two bowls were placed 150 cm away from the entrance and 280 cm apart from each other. The experimenter baited one of the two bowls surreptitiously with a piece of food. After baiting, the test subject was allowed to enter the test area from the starting area. Subjects received one of two different test conditions:

a) distal dynamic-sustained pointing kneeling (DDS-K) - The experimenter kneeled between the two bowls and as soon as the subject entered the corridor, he pointed and turned his head towards the baited bowl until the subject made a choice. The distance between the tip of the index finger and the baited bowl was about 80 cm.
b) distal momentary pointing kneeling (DM-K) - The experimenter kneeled between the two bowls and as soon as the subject entered the corridor, he pointed and turned his head towards the baited bowl for about one second or as long as the subject was still in the corridor. Pigs never entered the test area while the gesture was still being administered. The distance between the tip of the index finger and the baited bowl was about 80 cm.

Subjects received two test sessions of 20 trials each (10 trials for every condition in each session) with a total of 20 trials of each condition. Detailed methods and data are published in Nawroth et al. (2014).

Experiment 3

Fourteen domestic pigs (6 males, 8 females, 9 weeks) participated in this experiment. The experimenter baited one of the two bowls surreptitiously with a piece of food. After baiting, the test subject was allowed to enter the test area from the starting area. Subjects received one of two different test conditions:

a) kneeling behind correct location (behind) - The experimenter kneeled behind the baited bowl and remained there without moving, looking straight at the entrance.

b) pointing from incorrect location (incorrect) - The experimenter kneeled behind the non-baited bowl and as soon as the subject entered the corridor, he pointed and turned his head towards the baited bowl until the subject made a choice. The distance between the tip of the index finger and the baited bowl was about 80 cm. The tip of the index finger was always closer to the incorrect bowl than to the correct one.

Subjects received two test sessions of 20 trials each (10 trials for every condition in each session) with a total of 20 trials of each condition. Detailed methods and data are published in Nawroth et al. (2014).

Experiment 4

Thirteen domestic pigs (6 males, 7 females, 10 weeks) participated in this experiment. The experimenter baited one of the two bowls surreptitiously with a piece of food. After baiting, the test subject was allowed to enter the test area from the starting area. Subjects received one of three different test conditions:

a) proximal dynamic-sustained pointing (PSD) - The experimenter kneeled between the two bowls and as soon as the subject entered the corridor, he pointed towards the baited bowl until the subject made a choice, but remained looking straight forward. The distance between the tip of the index finger and the baited bowl was about 30 cm.

b) body orientation (body) - The experimenter was kneeled between the two bowls and as soon as the subject entered the corridor, he oriented his body and head towards the baited bowl until the subject made a choice. The distance between the experimenter´s face and the baited bowl was about 100 cm. The distance between the experimenter´s knee and the baited bowl was about 70 cm, whereas the distance to the incorrect bowl was about 75 cm.

c) head orientation (head) - The experimenter was kneeled between the two bowls and as soon as the subject entered the corridor, he turned his head towards the baited bowl until the subject made a choice. The distance between the experimenter´s face and the baited bowl was about 100 cm.

Subjects received three test sessions of 20 trials each (two trials for every condition in each session) with a total of 18 trials of each condition. In a fourth session, six test trials (two for each condition) were administered, resulting in a total of 20 trials for each condition. Detailed methods and data are published in Nawroth et al. (2014).

Reference

Nawroth, C., Ebersbach, M., von Borell, E. (2014) Juvenile domestic pigs (Sus scrofa domestica) use human-given cues in an object choice task. Animal Cognition 17(3): 701–713

27. SI Nawroth et al. 2013-14a

Author's name and affiliation:
Nawroth Christian[1], von Borell Eberhard[2], Jan Langbein[1]

[1]Leibniz Institute for Farm Animal Biology, Institute of Behavioural Physiology, Dummerstorf, Germany
[2]University of Halle-Wittenberg, Institute of Agricultural and Nutritional Sciences, Halle, Germany

Author's contribution:
CN, EvB and JL planned the study. CN collected the data. CN and JL analysed the data. CN wrote the SI methods.

Phil trans authors requested:
Jan Langbein

Data sharing (full dataset or summary data):
Summary data will be shared.

*Methods.*We collected data on the use of indirect visual information in domestic goats (*Capra hirca*) during 3 different experiments carried out in Dummerstorf (Germany):

- *Exp1: Use of indirect visual information* April 2013
- *Exp1: Use of indirect visual information, additional control* April 2013
- *Exp3*: *Use of indirect acoustic information* May 2014

For all experiments, data were collected using an object choice task. For training and testing, the goats were separated in a compartment adjacent next to their home pen (150 cm x 125 cm). An experimenter was seated in another compartment, separated from the test animal by a grate, leaving the subjects several spaces within the grate where they could indicate a choice. A sliding board (60 cm x 25 cm) was placed in front of the grate. Two dark brown bowls (diameter: 14 cm) were placed on the board with a distance of 35 cm. Two dark brown cups (diameter: 11 cm; height: 10 cm) were used to cover the bowls. The distance between the bowls and the test subject was approximately 30 cm. Habituation of subjects to the test arena and detailed training procedures are described in Nawroth et al. (2014).

Experiment 1: Use of indirect visual information
Eleven Nigerian dwarf goats (all female, 3-4.5 years) participated in this experiment. The experimenter baited one of the two bowls surreptitiously with a piece of food and covered both bowls with the corresponding cups. The experimenter than placed both bowls and cups on the sliding board. Subjects received one of four different test conditions:

1) both – the experimenter lifted both cups simultaneously for approximately 5 seconds, giving full information of the location of the reward to the subject;

2) direct - the experimenter lifted the baited cup for approximately 5 seconds while simultaneously touching the non-baited cup, giving only direct information of the location of the reward to the subject;

3) indirect – the experimenter lifted the non-baited cup for approximately 5 seconds while simultaneously touching the baited cup, giving only indirect information of the location of the reward to the subject;

4) control – the experimenter touched both cups simultaneously without lifting them for approximately 5 seconds, giving no information of the location of the reward to the subject.

Subjects received ten test sessions of eight trials each (two trials for every condition in each session) with a total of 20 trials of each condition. Detailed methods and data are published in Nawroth et al. (2014).

Experiment 2: Use of indirect visual information, additional control
Eleven Nigerian dwarf goats (all female, 3-4.5 years) participated in this experiment.

The procedure was the same as in Experiment 1, except that underneath the bigger outer cups two smaller inner cups, either transparent or opaque, were positioned. In all conditions, both outer cups were lifted. To reproduce the four informational levels described in Experiment 1, the inner cups were either opaque or transparent. Subjects received one of four different test conditions:

1) both – the experimenter lifted both outer cups simultaneously for approximately 5 seconds; both inner cups were transparent

2) direct - the experimenter lifted both outer cups simultaneously for approximately 5 seconds; the inner baited cup was transparent, the inner non-baited cup was opaque

3) indirect – the experimenter lifted both outer cups simultaneously for approximately 5 seconds; the inner baited cup was opaque, the inner non-baited cup was transparent

4) control – the experimenter lifted both outer cups simultaneously for approximately 5 seconds; both inner cups were opaque

Subjects received ten test sessions of eight trials each (two trials for every condition in each session) with a total of 20 trials of each condition. Detailed methods and data are published in Nawroth et al. (2014).

Experiment 3: Use of indirect acoustic information
Six Nigerian dwarf goats (all female, 4-5.5 years) participated in this experiment. The general methods were similar to Experiment 1, with a few exceptions: the experimenter baited one of the two cups (positioned upside-down) surreptitiously with a piece of food. The experimenter than placed both cups on the sliding board. Subjects received one of four different test conditions:

1) both – the experimenter lifted and shook both cups simultaneously for approximately 3 seconds, giving full information of the location of the reward to the subject;

2) direct - the experimenter lifted both cups and shook the baited cup for approximately 3 seconds, giving only direct information of the location of the reward to the subject;

3) indirect – the experimenter lifted both cups and shook the non-baited cup for approximately 3 seconds, giving only indirect information of the location of the reward to the subject;

4) control – the experimenter lifted both cups simultaneously, without shaking them, for approximately 3 seconds, giving no information of the location of the reward to the subject.

Subjects received six test sessions of 24 trials each (four trials for every condition in each session) with a total of 24 trials of each condition. The data are unpublished.

Reference

Nawroth, C., von Borell, E., Langbein, J. (2014) Exclusion performance in dwarf goats (*Capra aegagrus hircus*) and sheep (*Ovis orientalis aries*). PLoS ONE 9(4): e93534.

28. SI Nawroth et al. 2015

Author's name and affiliation:
Nawroth Christian[1], von Borell Eberhard[2], Jan Langbein[1]

[1]Leibniz Institute for Farm Animal Biology, Institute of Behavioural Physiology, Dummerstorf, Germany
[2]University of Halle-Wittenberg, Institute of Agricultural and Nutritional Sciences, Halle, Germany

Author's contribution:
CN, EvB and JL planned the study. CN collected the data. CN and JL analysed the data. CN wrote the SI methods.

Phil trans authors requested:
Jan Langbein

Data sharing (full dataset or summary data):
Summary data will be shared.

Methods. We collected data on the use of human-given cues in domestic goats (*Capra hircus*) during 3 different experiments carried out in Dummerstorf (Germany) over a period of 3 weeks. Each experiment ran over approximately 1 week.

For all experiments, data were collected using an object choice task. For training and testing, the goats were separated in a compartment adjacent next to their home pen (150 cm x 125 cm). An experimenter was seated in another compartment, separated from the test animal by a grate, leaving the subjects several spaces within the grate where they could indicate a choice (Figure 1). A sliding board (60 cm x 25 cm) was placed in front of the grate. Two cups (various diameters and colours for each experiment) were placed on the board with a distance of 35 cm. The distance between the cups and the test subject was approximately 30 cm. Habituation of subjects to the test arena and detailed training procedures are described in Nawroth et al. (2014).

Experiment 1: cups upside-down
Ten Nigerian dwarf goats (all female, 4-5.5 years) participated in the experiment. The experimenter put a reward on either the left or right side of the sliding board for 2 seconds before both sides were covered with a cup (brown: Ø 9 cm). The experimenter than moved the left cup to the right side and the right cup to the left side of the board so that the cups crossed their path in the middle. After the transposition, the experimenter waited for 2 seconds until the sliding board was pushed towards the grating, allowing the subjects to make a choice. Each subject received one test session that consisted of twelve test trials. Detailed methods and data are published in Nawroth et al. (2015).

Experiment 2: different coloured cups
Nine Nigerian dwarf goats (all female, 4-5.5 years) participated in the experiment. The test procedure was similar to Experiment 1, except that two cups differing in colour and size were used for the transposition task (dark brown: Ø 11 cm; white: Ø 9 cm). Each subject received one test session that consisted of twelve test trials. Detailed methods and data are published in Nawroth et al. (2015).

Experiment 3: same coloured cups
Ten Nigerian dwarf goats (all female, 4-5.5 years) participated in the experiment. The test procedure was similar to Experiment 2, except that both of the cups were identical in shape and colour (dark brown: Ø: 11 cm). Each subject received one test session that consisted of twelve test trials. Detailed methods and data are published in Nawroth et al. (2015).

References

Nawroth, C., von Borell, E., Langbein, J. (2014) Exclusion performance in dwarf goats (*Capra aegagrus hircus*) and sheep (*Ovis orientalis aries*). PLoS ONE 9(4): e93534.

Nawroth, C., von Borell, E., Langbein, J. (2015) Object permanence in the dwarf goat (*Capra aegagrus hircus*): Perseveration errors and tracking of complex movements of hidden objects. Applied Animal Behaviour Science 167: 20–26.

29. SI Nawroth et al 2013-14b

Author's name and affiliation:
Nawroth Christian[1], von Borell Eberhard[2], Jan Langbein[1]

[1]Leibniz Institute for Farm Animal Biology, Institute of Behavioural Physiology, Dummerstorf, Germany
[2]University of Halle-Wittenberg, Institute of Agricultural and Nutritional Sciences, Halle, Germany

Author's contribution:
CN, EvB and JL planned the study. CN collected the data. CN and JL analysed the data. CN wrote the SI methods.

Phil trans authors requested:
Jan Langbein

Data sharing (full dataset or summary data):
Summary data will be shared.

Methods. We collected data on the use of human-given cues in domestic goats (*Capra hircus*) during 2 different experiments carried out in Dummerstorf (Germany). Each experiment ran over 1 week:

- *Exp1:* April 2013
- *Exp2*: May 2014

For all experiments, data were collected using an object choice task. For training and testing, the goats were separated in a compartment adjacent next to their home pen (150 cm x 125 cm). An experimenter was seated in another compartment, separated from the test animal by a grate, leaving the subjects several spaces within the grate where they could indicate a choice. A sliding board (60 cm x 25 cm) was placed in front of the grate. Two dark brown bowls (diameter: 14 cm) were placed on the board with a distance of 35 cm. Two dark brown cups (diameter: 11 cm; height: 10 cm) were used to cover the bowls. The distance between the bowls and the test subject was approximately 30 cm. Habituation of subjects to the test arena and detailed training procedures are described in Nawroth et al. (2014).

Experiment 1

Eleven Nigerian dwarf goats (all female, 3-4.5 years) participated in this experiment. The experimenter baited one of the two bowls surreptitiously with a piece of food and covered both bowls with the corresponding cups. The experimenter than placed both bowls and cups on the sliding board. Subjects received one of four different test conditions:

1) touch - the experimenter touched the baited cup for 3 seconds

2) point - the experimenter pointed at the baited cup for 3 seconds (dynamic sustained pointing, distance to baited cup: 5 cm)

3) head only - the experimenter oriented his head towards the baited cup for 3 seconds

4) control - the experimenter remained motionless for 3 seconds

Subjects received six test sessions of 14 trials each (four trials for each test condition and two trials for the control condition in each session) with a total of 24 trials per test condition and twelve trials for the control condition. Detailed methods and data are published in Nawroth et al. (2015).

Experiment 2

Ten Nigerian dwarf goats (all female, 4-5.5 years) participated in this experiment. The general setup was similar to Experiment 1. Subjects received one of two different test conditions:

1) crosspoint near - the experimenter pointed at the baited cup with his contralateral arm for 3 seconds, (dynamic sustained pointing, distance to baited cup: 10 cm; the experimenter's arm did protrude the upper torso)

2) crosspoint far - the experimenter pointed at the baited cup with his contralateral arm for 3 seconds (dynamic sustained pointing, distance to baited cup: 40 cm, the experimenter's arm did not protrude the upper torso)

Subjects received one test sessions of 12 trials with a total of 6 trials of each condition. The data are unpublished.

References

Nawroth, C., von Borell, E., Langbein, J. (2014) Exclusion performance in dwarf goats (*Capra aegagrus hircus*) and sheep (*Ovis orientalis aries*). PLoS ONE 9(4): e93534.

Nawroth, C., von Borell, E., Langbein, J. (2015) 'Goats that stare at men' – Dwarf goats alter their behaviour in response to human head orientation but do not spontaneously use head direction as a cue in a food-related context. Animal Cognition 18(1): 65–73.

30. SI Oesterwind (unpublished data)

Serial reversal-learning in Nigerian dwarf goats

Author's name and affiliation:
Susann Oesterwind[1], Jan Langbein[2], Katrin Siebert[2], Antonine Finkemeier[1]

[1]Behavioural Sciences, Faculty of Agricultural and Environmental Sciences, University of Rostock, Germany
[2]Leibniz Institute for Farm Animal Biology (FBN), Institute of Behavioural Physiology, Dummerstorf, Germany

Author's contribution:
SO, JL planned the study. SO, KS and AF collected the data. SO, KS and JL analysed the data. JL wrote the SI methods.

Phil trans authors requested:
Susann Oesterwind, Jan Langbein

Data sharing (full dataset or summary data):
Summary data will be shared.

Methods. We collected data on serial reversal-learning of Nigerian dwarf goats (Capra hircus, all females) using a four-choice visual-discrimination paradigm. Experiments were conducted at the Leibniz Institute for Farm Animal Biology (FBN) in Dummerstorf (Germany). We tested a total of 108 goats in four test runs over a period of two years. During all phases of the experiment, goats were housed in groups of up to 10 animals in pens with straw as bedding, two times concentrate per day and hay ad libitum. Data were collected using a fully automated learning device (LD) developed at the FBN (Figure 1A).

The LD was integrated into the pen of the goats, and the animals had access to the device 24 hours a day. Only one goat could act at the LD at a time, however, animals had non-restricted access to device. The LD has already been described in detail [1]. Four black symbols were presented on a white computer screen. One symbol was rewarded (S+), while the three other symbols were used as distractors (S−). Two series of the 24 possible image combinations were mixed to a pseudorandom pattern series of 48 patterns in a row. We ensured that S+ was not on identical positions in consecutive trials. To choose a symbol, the goat had to press a buttons placed next to the symbol. After choosing the rewarded symbol, the goat received a small amount of drinking water (30 ml). All individual visits at the LD and all button presses were recorded automatically. The controlling software ensured that side preferences were counteracted at any time.

Reversal learning experiment

Shaping

After weaning (at the age of six weeks), goats were grouped together in a pen equipped with a simplified LD. They were trained stepwise over a six-week period to press different buttons to get drinking water as a reward [2]. After they reliably pressed changing buttons every day for a reward, the LD was introduced. After one week with a white screen, where all four 4 buttons were rewarded, two first 4-choice discrimination problems were trained, each of which ran for 14 days. From a former study, we know dwarf goats stabilize learning performance after

training of two to three consecutive discrimination problems [3].

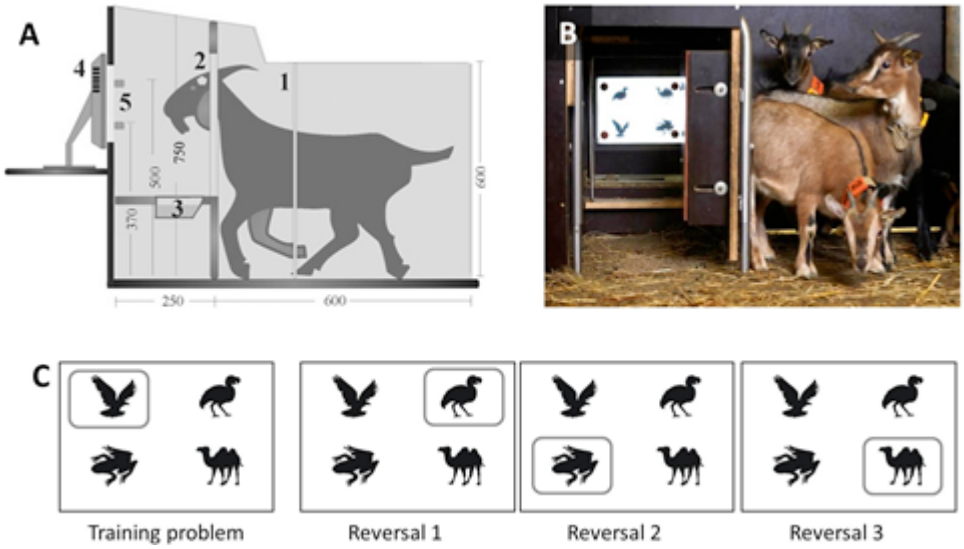


Figure 1. A) Sketch and measurements (mm) of the LD: 1 = aerial for individual identification, 2 = head gate, 3 = reward serving bowl, 4 = computer screen, 5 = four buttons, each related to one symbol on the screen. B) The LD integrated into the goat pen. C) The originally trained pattern combination and the three reversal problems. The rewarded symbol in each problem is framed.

Training and reversal tests

The animals were trained to a new 4-choice discrimination problem (Figure 1) for 14 days. After evaluating the steepness of the learning curve of the training problem, the three reversal problems were presented for seven days each. Each discrimination problem was trained directly after the previous one. We calculated for each animal the number of trials required to achieve the learning criterion in the training problem and in the three reversal problems. The learning criterion was defined as 46 % of correct choices in at least two consecutive sequences of 20 trials ($P < 0.05$; binomial test; $N = 20$; $P0 = 0.25$). We selected data from the 75 animals that have reached the test criterion in the training problem and at least one reversal problem within the given time period. The data are unpublished.

References

[1] Langbein J, Siebert K, Nürnberg G, Manteuffel G. 2007 The impact of acoustical secondary reinforcement during shape discrimination learning of dwarf goats (Capra hircus). Appl. Anim. Behav. Sci. 103, 35-44. (doi:10.1016/j.applanim.2006.04.019).

[2] Langbein J, Nurnberg G, Manteuffel G. 2004 Visual discrimination learning in dwarf goats and associated changes in heart rate and heart rate variability. Physiol. Behav 82, 601-609. (doi:10.1016/j.physbeh.2004.05.007).

[3] Langbein J, Siebert K, Nurnberg G, Manteuffel G. 2007 The impact of acoustical secondary reinforcement during shape discrimination learning of dwarf goats (Capra hircus). J. Comp. Psychol. 121, 447-456. (doi:10.1037/0735-7036.121.4.447).

31. SI Sorato & Lovlie

Author's name and affiliation:
Enrico Sorato[1], Hanne Løvlie[1]

[1]IFM Biology, Linköping University, 58283 Linköping, Sweden

Author contribution
ES and HL collected data, ES coordinated data handling and ran preliminary analysis, HL coordinated the study. HL wrote the SI methods with input from ES.

Methods. Between May-December 2016, a total of 113 red junglefowl (*Gallus gallus*) from a captive population maintained at Linköping University were assayed in a discrimination learning task (chicks: age 3-6 days, $n_{males} = 35$, $n_{females} = 37$; adults: > 6 months, $n_{males} = 23$, $n_{females} = 46$), and a reversal learning task (age 5-7 days, $n_{males} = 33$, $n_{females} = 34$). Of these, 28 individuals were tested both as chicks and when adult ($n_{males} = 8$; $n_{females} = 20$), 44 were tested only as chicks ($n_{males} = 26$; $n_{females} = 18$), and 41 were assayed only when adult ($n_{males} = 16$; $n_{females} = 25$). 23 adult males and 6 adult females were not motivated in the discriminant learning task and were therefore excluded from analysis, while all chicks reached our learning criterion. The relative low number of males assayed as adults was due to a general lack of interest in food and weariness for human presence.

Birds were housed in groups and tested singly during daytime (8-18, lights on 7-19 local time), further details of testing procedures are provided elsewhere (Zidar et al, 2017a,b; Zidar et al, submitted; Sorato et al, 2018, this issue). In brief, 'learning speed' was quantified for each task as the total number of trials needed to discriminate between two colour cues (each consisting of a coloured card, 9 x 9 cm, and a bowl in the same colour, 5Ø x 3H cm for chicks, 5Ø x 7H cm for adults), by meeting a learning criterion of 6 consecutive correct choices. For chick discriminative learning, a black cue was rewarded with a piece of mealworm, while a white cue was unrewarded. For chick reversal learning and adult discriminant learning, the white cue was rewarded, while the black cue was unrewarded. Performance of adults previously exposed (age 3-6 days) to the learning tasks did not differ from the performance of 'naïve' adult individuals, ruling out long-lasting memory or other effects of previous exposure to the cognitive assays. Tasks were carried out in a test arena (L x W x H, chicks: 46 x 36 x 18 cm, adults: 100 x 60 x 80 cm), wherein the bird was placed at the opposite end of the two cues, and trials lasted from when the bird was placed in the arena until it made a choice (i.e. when it had its head within 2cm of the cue). An individual was allowed as many trials as it could perform within a session (for chicks a session lasted 15 min; for adults 30min due to longer motivation span than chicks). If further trials were needed to reach the learning criterion, a new session was started after ≥ 1 hr (with the bird back into its group in the meantime). Once a chick had reached the learning criterion for discriminative learning, it was exposed to the reversal learning task in a new session. If > 7 hours had passed since the final discriminative learning session, the chick was exposed to a 'refresh' session in which it had to reach again the learning criterion for discriminative learning before continuing to the reversal learning test. This was done to ensure that the association between the previously learned cue and the reward was still salient before performing reversal learning. Black and white cues were presented left-right in a predetermined, pseudorandom order, as to avoid any side-preference effect.

References

Sorato E, Zidar J, Garnham L, Wilson A, & Løvlie H. Heritability and co-variation among cognitive traits in the red junglefowl. Phil. Trans. Roy. Soc. B., this issue

Zidar J, Balogh A, Favati A, Jensen P, Leimar O, Sorato E, & Løvlie H. The relationship between learning speed and personality is age- and task-dependent in the red junglefowl. Submitted.

Zidar J, Sorato E, Malmqvist A-M, Jansson E, Rosher C, Jensen P, Favati A, & Løvlie H (2017a) Early experience affect adult personality in the red junglefowl: a role for cognitive stimulation? Behav. Proc. Special issue on individual variation in cognition and personality. 134: 78-86.

Zidar J, Balogh A, Favati A, Jensen P, Leimar O, & Løvlie H (2017b) A comparison of animal personality and coping styles in red junglefowl. Anim. Behav. 130: 209-220. doi: 10.1016/j.anbehav.2017.06.024.

Data availability
Data will be submitted to Dryad.

32. SI van Horik JO & Emery NJ

Author's name and affiliation:
Jayden van Horik and Nathan Emery
Queen Mary University of London, UK

Author's contribution:
JOvH and NJE planned the study. JOvH collected the data, analysed the data and wrote the SI methods.

Phil trans authors requested:
Jayden van Horik

Data sharing (full dataset or summary data):
Data will be shared.

The delay between each reversal was approximately 7 days

Parrots Methods: serial reversal learning

General methods

Subjects and Housing

Four red-shouldered macaws (*Diopsittaca nobilis*): No.2, No.4, No.5, and No.8, and four black-headed caiques (*Pionites melanocephala*): Green, Gold, Purple, and Red, participated in this study (hereafter macaws and caiques). All subjects were male, with the exception of one female macaw (No.4). All subjects were hand-reared, approximately two years old when tested. Each species were housed in separate indoor aviaries ($2m^3$). None of the subjects had prior experience with serial reversal learning tasks, but they were experienced with a number of tasks employing object manipulation, including removing food hidden under lids or cups. Both species were raised under identical conditions and provided with equal experiences. Food and water were provided *ad libitum* and subjects' participation was voluntary.

Apparatus and Training

During training trials the apparatus, a symmetrical wooden base (28 cm x 7 cm) with two food wells (1.5 cm dia) separated by 12 cm, was presented to subjects. One food-well contained a reward of crushed Lafeber Nutri-Berries, while the other well remained empty. After subjects fed from the apparatus without hesitation, two orange 6 cm diameter plastic lids were fixed to hinges to conceal the contents of the wells. Again, only one well was baited. The location of the baited well was pseudorandomised accross training trials so that it did not occur on the same side over more than two consecutive trials. This procedure attempted to control for the formation of side biases and facilitate subjects' searching behaviours. To proceed to test, subjects were required to retrieve the concealed food by opening the lids at least ten times in one 10min session.

Procedure

Subjects were not food deprived, although testing was conducted in the morning prior to their regular feeding schedule. Each subject was provided with one session of 10 trials per day. The presentation of rewarded and un-rewarded coloured lids, coloured either blue or green, was counterbalanced across subjects. To prevent the development of side biases, the position of the lids (i.e. left or right hand side presentation) was pseudorandomised within sessions so that the lids did not occur on the same side for more than two consecutive trials.

If subjects reached a predetermined criterion of seven consecutive correct trials in one block of 10 trials (significant according to a binomial test with a probability of choosing either side set at 0.5), they were immediately presented with one block of 10 trials with reversed contingencies (i.e. S+ becomes S– and vice versa). To avoid satiation and encourage motivation to interact with the apparatus, subjects were only presented with one post-reversal block per day. Hence, subjects could only receive a maximum of two consecutive blocks of 10 trials per day. There were no occurrences where subjects reached criterion again during their first post-reversal block. Each subject was presented with as many blocks as required to reach eight serial reversals.

Each subject was tested individually in a familiar enclosure ($2m^3$) where they were visually isolated from all other subjects. During testing days, all subjects participated in the experiment in a randomised order. Subjects were familiar with being handled by the experimenter and were transferred to the experimental cage by hand. Daily trials typically began at 08:30 and ceased around 13:00 although duration of each testing session, and the corresponding inter-trial intervals, varied depending on the subjects motivation to interact with the apparatus. The duration of a typical testing session was between 15-20 minutes per bird. During testing trials, the experimenter attempted to avoid providing subjects with any inadvertent cues to the location of the concealed reward by holding and presenting the apparatus in a symmetrical fashion and then placing his hands behind his back and looking only at the centre of the apparatus. Subjects were only allowed to upturn one lid per trial and were considered to have made a correct choice if they chose the baited lid. Hence, if subjects upturned the correct lid, they were allowed to retrieve the food reward. However, if subjects upturned the un-baited lid, then the apparatus was immediately removed. If subjects failed to upturn the baited lid on one trial, the succeeding trials followed the predetermined pseudorandomised order. The apparatus was re-baited out of view of the subject. Subjects that chose the same side over six consecutive trials in one block were considered to have developed a side bias. To correct for side biases, we presented the baited lid on the non-preferred side until the subject chose the baited side for two consecutive trials. Trials then reverted to the original pseudorandomised configuration. All trials, including side-bias-corrected and non-corrected trials were included in the subsequent analyses. We recorded all trials with a digital camcorder (JVC Everio, Model No. GZ-MG645BEK, Malaysia) and scored the number of number of trials and the number of errors to reach criterion for the initial colour association and for each subsequent reversal.

33. SI van Horik, JO & Madden, JR 2016

Author's name and affiliation:
Jayden van Horik and Joah R Madden
Department of Psychology, University of Exeter, UK

Author's contribution:
JOvH and JRM planned the study. JOvH collected the data, analysed the data and wrote the SI methods.

Phil trans authors requested:
Jayden van Horik

Data sharing (full dataset or summary data):
Data will be shared.

The approximate delay between each discrimination task was 5 days

Methods. We reared 200 day-old pheasant chicks (*Phasianus colchicus*) in groups of 50 in four replicated enclosures and between 28 May 2015 and 29 July 2015. All subjects were individually marked using numbered wing tags, fed on commercial pheasant feed supplemented with wild bird seed (~5%) and supplied with water *ad libitum.* Birds were housed in 2m x 2m heated huts for the first 2 weeks of life. They had access to unheated but covered outdoor runs of 1m x 4m for the next week and for the final seven weeks of rearing had access to 4m x 12m outdoor runs. All birds were tested with a battery of psychometric tests (including those detailed in this study) from 10 days old, with equal exposure in a fixed order to all tasks (van Horik, Langley, Whiteside, & Madden, 2016). During test sessions subjects could enter the experimental chamber (75cm x 75cm) at will, where they were tested individually while visually isolated from other test subjects. Morphometrics (mass, tarsus length) were taken and sex confirmed by plumage features at ten weeks old when testing ceased.

Procedures
Subjects were initially trained, using shaping procedures, to peck through a layer of crepe paper and retrieve a mealworm reward concealed in a well van Horik et al., 2016). During testing, subjects were presented with two colour discrimination tasks (Green/Blue: 28 – 30 June 2015; Yellow/Pink: 06 – 8 July 2015) involving an acquisition learning phase and a reversal learning phase. Each task required subjects to discriminate between two wells in which the contents were concealed by a layer of crepe paper. One well contained a mealworm reward while the other well was made inaccessible by covering it with hard black card placed under the crepe paper, which could not be pecked through. Rather than leaving the well empty, we considered the black card to provide a more salient cue of an incorrect choice. Each well was encircled by one of two colour cues. During the Green/Blue acquisition phase, the rewarded well was associated with a green cue and the unrewarded well was associated with a blue cue. During the Yellow/Pink acquisition phase the rewarded well was associated with a yellow cue and the unrewarded well associated with a pink cue.

Each subject was presented with five sessions of each discrimination pair. Subjects received two sessions per day, one in the morning and one in the afternoon, making ten binary choices (hereafter 'trials') in each session. Therefore, each bird received a total of 50 trials. A correct

choice was scored if subjects first pecked into a rewarded well and an incorrect choice was scored if subjects first pecked into an unrewarded well. If the bird made a correct choice, it was allowed to eat the reward. If the bird made a wrong choice, the pair of wells was removed and replaced with a new pair. The location of the rewarded well was pseudorandomised across trials, and did not occur on the same side for more than three consecutive trials.

Fitting individual Learning Curves

We used learning curves to summarize individual performance across trials. Four learning curves were generated for each individual, one for each of the four different tasks. Learning curves were generated for 187 individuals that completed all 50 trials in at least one of the four discrimination tasks. However, our analyses across the different Acquisition and Reversal tasks were restricted to only 111 individuals that completed all trials on all four tasks (n = 59 males, n = 49 females, and three individuals for which we did not have sex or body condition measures). The coefficients describing learning curves were generated from whether or not a given subject made a correct or incorrect choice per trial, after fitting a sigmoid curve to the binary choice data using R (R Development Core Team, 2014). For our learning criteria, we used the predicted trial number when the curve crossed a line indicating that there was an 80% probability of the bird making a correct choice. We derived this measure by solving the equation $X = (-\ln 0.25 - b_0)/b_1$, where b_1 is the slope of the learning curve, and b_0 is the intercept. Learning curves accounted for individuals with a strong positive bias, as these birds showed poor improvement in performance. Our derived trial numbers were log transformed prior to analysis to improve normality of the data.

References

R Development Core Team. (2014). R: a language and environment for statistical computing. Vienna Austria: R Foundation for Statistical Computing.

van Horik, J. O., Langley, E. J. G., Whiteside, M. A., & Madden, J. R. (2016). Differential participation in cognitive tests is driven by personality, sex, body condition and experience. *Behavioural Processes*, 1–9. https://doi.org/10.1016/j.beproc.2016.07.001

34. SI Wallis et al. 2016

Author's name and affiliation:
Lisa J. Wallis [1,2], Zsófia Virányi [2], Corsin A. Müller [2], Samuel Serisier [3], Ludwig Huber [2], and Friederike Range [2]

[1] Department of Ethology, Eötvös Loránd University, Budapest, Hungary
[2] Clever Dog Lab, Messerli Research Institute, University of Veterinary Medicine Vienna, Medical University of Vienna, University of Vienna, Vienna, Austria
[3] Royal Canin Research Center, Aimargues, France

Author's contribution: LH, ZV and FR planned the study. LW collected and analysed the data, and wrote the SI methods. LW wrote the paper with ZV CM SS LH and FR.

Phil trans authors requested:
Lisa J. Wallis and Ludwig Huber

Data sharing (full dataset or summary data):
Full data set available

Methods. Reduced methods are presented here of the three tests that were used to measure repeatability in the current study. Further details can be obtained from the original published article and supplementary materials (see reference Wallis et al. 2016). Figures and excerpts from the methods from the open access publication "Aging effects on discrimination learning, logical reasoning and memory in pet dogs" are reproduced here with permission from the authors. https://link.springer.com/article/10.1007/s11357-015-9866-x

Subjects

One hundred and five pet dogs ranging in age from 5 months to 13 years and 10 months were recruited to participate in the study (Table 1). Data from ten extra dogs were added after publication of Wallis et al. 2016. All dogs were from one breed, the Border Collie, in order to exclude the effects of different developmental and aging speeds of different breeds.

Table 1: Age, sex and neuter status of subjects, and number of individuals that completed the geometric forms, underwater drawings, and clip art pictures discriminations.

Age group	Life stage	Age in years	Male (neutered)	Female (neutered)	Total	Geometric forms	Underwater drawings	Clip art pictures
Group 1	Late puppyhood	0.33 to 1	7 (0)	15 (1)	22	22	20	19
Group 2	Adolescence	> 1 – 2	10 (1)	11 (2)	21	21	21	20
Group 3	Early adulthood	> 2 – 3	9 (3)	15 (5)	24	24	21	20
Group 4	Middle age	> 3 – 6	6 (2)	13 (6)	19	19	18	14
Group 5	Late adulthood	> 6	7 (5)	12 (12)	19	19	13	14
Total			37 (11)	68 (26)	105	105	93	87

Apparatus

Testing was carried out at the Clever Dog Lab in Vienna, Austria. The test apparatus consisted of a closed rectangular box containing the food pellet dispenser and an adjacent testing niche. Dogs were tested in the testing niche, which allowed subjects to operate the touchscreen whilst avoiding potential distractions from the side or above, thus minimising human influence on the dogs' performance. A small hole beneath the touchscreen allowed commercial dog food pellets to be automatically dispensed in order to administer reinforcement for correct choices.

Procedure

The touchscreen training and testing procedures consisted of two pre-training steps: an approach training and a simple geometric form discrimination, and two tasks: a 'categorical' discrimination (underwater photographs and drawings), and a clipart picture discrimination.

Touchscreen pre-training (for details of approach training please refer to Wallis et al. 2016)

Geometric form discrimination

Subjects were presented with a square and a circle side by side. Both stimuli varied in colour between trials (red, yellow or blue, Figure 1a). The dogs were assigned to two groups balanced for age group and sex. Group 'square' was rewarded for touching the square; group 'circle' was rewarded for touching the circle. The two shapes were presented simultaneously on a black background in fixed positions on the screen (at the animal's eye level, one appearing left of the middle, and the other right, Figure 1a). Each trial was composed of one positive stimulus (S+) and one negative stimulus (S−), which were positioned randomly from trial to trial (left/right). Each session consisted of 30 trials. When the positive stimulus was selected, both stimuli disappeared, a short tone was emitted by the computer, and a food reward was provided. If the wrong stimulus was touched (S−), both stimuli disappeared, a short buzz sounded, and a red screen was presented for 3 seconds. In this case, a correction trial was immediately initiated:

the stimuli of the previous trial were presented again in the same positions. A correct choice terminated the trial and resulted in reward and presentation of a new trial. After each trial (except correction trials), an inter-trial interval of 2 s was initiated (an empty black background was presented). The learning criterion was set at ≥20 correct first choices in 30 trials (66.7 %) in four out of five consecutive sessions.

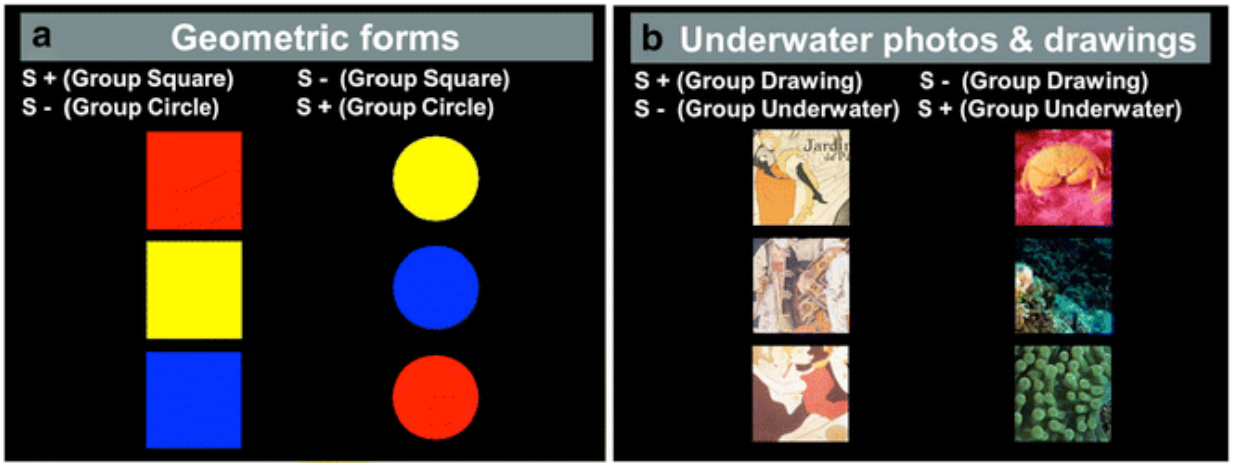


Figure 1: Training stimuli for the a) geometric form and b) underwater photo and drawing discriminations

Touchscreen testing

Task 1: Underwater photos and drawings discrimination

Once the criterion for the geometric form task was reached, the dogs were transferred to a second discrimination training, involving three underwater photographs, which had to be distinguished from three drawings (two of which were taken from posters by Toulouse-Lautrec; Figure 1b). The dogs were assigned to two groups balanced for age group and sex. Group 'drawing' was rewarded for touching the drawing and group 'underwater' was rewarded for touching the underwater photograph. In each trial, one of the three S+ was randomly coupled side by side with one of the three S−. The procedure and learning criterion were the same as for the geometric form discrimination.

Task 2: Clip art picture discrimination

Once the dogs had completed the underwater photos and drawing discrimination, they then moved onto the clipart picture discrimination. Dogs were again split into two groups (Group 'A' and Group 'B') balanced for age group and sex. The dogs were trained to discriminate four S+ and four S− stimuli (Figure 2), this time presented on a white background. The stimuli were coloured clip art pictures obtained from the internet and were grouped within the two sets by avoiding similarities in colour, form or function. Each session consisted of 32 trials and contained each of the 16 possible S+/S− pairings twice per session. All dogs were required to reach a learning criterion of ≥28 correct first choices (87.5 %) in two consecutive sessions.

Figure 2: Reason by exclusion training stimuli

Data Summary Table

Dog were allocated into age groups based on their age in months on the first day of training for the geometric forms discrimination. However, the age in months of the dogs is presented for all three discriminations (geometric forms, underwater photos and drawings, and clip art picture discriminations), to allow the calculation of the time between the three discriminations (which could be different for each dog depending on how long they took to complete the learning criteria). The stimulus group that the dog was allocated to is presented for all three tests (geometric forms (Circle/Square: three different colours), underwater photos and drawings (three different underwater photos, and three different drawings), and clip art picture discriminations (Group A/B: four positive and four negative)). We created a new variable that described the difficulty of the stimulus group, which consisted of either "Hard" or "easy".

For the three discriminations the number of sessions to criteria is used as a measure of the dogs learning ability.

Sessions to criterion: The total number of sessions needed for the dogs to reach the criterion. For a few of the dogs which were tested using an early version of the software, it occasionally happened that the dog completed either more or less trials in a session than was allocated (and so it is possible that the sessions until criteria may not be a whole number).

Results

From the paper utilising a slightly smaller sample the following factors were found to be significant predictors in statistical models when the response variable was "Sessions to criterion":-
Geometric forms: Stimulus group
Underwater photos and drawings: Stimulus group and age in months
Clip art picture: Stimulus group, sex, and age in months
For details, please see Wallis et al. 2016 and supplementary materials.

Reference

Wallis, L. J., Virányi, Z., Müller, C. A., Serisier, S., Huber, L., & Range, F. (2016). Aging effects on discrimination learning, logical reasoning and memory in pet dogs. *Age*, *38*(1), 6.

35. SI Wilkison (unpublished)

Author's name and affiliation:
Emma Smith and Anna Wilkinson
School of Life Sciences
University of Lincoln, UK

Author contribution:
ES and AW designed the study and collected the data.

Phil trans authors requested:
Emma Smith and Anna Wilkinson

Data sharing (full dataset or summary data):
Full datasets will be shared.

Acknowledgements:
We would like to thank Sabine and Thomas Vinke for allowing us to use their animals and their help throughout the project.

Methods.
We ran a serial reversal learning task in the red-footed tortoise. This data is based on two experiments investigating visual discrimination and reversal learning in the red-footed tortoise (*Chelonoidis carbonaria*). In the first, animals were trained to discriminate between two sets of stimuli that varied in both colour and shape. In the second experiment, the contingencies of the trained stimuli were reversed four times, this was followed by a generalization test.

Data were collected from captive red-footed tortoises held in the University of Lincoln's cold-blooded cognition laboratory (n=4) and animals (n=3) kept in semi- free ranging captive conditions in Paraguay.

Experiment 1

General Procedure
Prior to the onset of the experiment, animals were habituated to a testing arena until they readily ate in the environment. Tortoises were then trained to approach a specific stimulus. Once they had learned to do so, so they were trained to make a discrimination using a two-alternative forced choice procedure.

Stimuli consisted of colored shapes and therefore differed in two dimensions. At the onset of each trial, stimuli were positioned 24cm apart at one end of the arena, the tortoise was then placed at the starting position which was equidistant between the two stimuli. A tortoise was considered to have made a choice when the animal was within 5cm of the stimulus and looking directly at it. If an animal chose the correct stimulus, then a small piece of favored food was delivered before it was removed from the arena. If an animal chose the incorrect stimulus, then it was removed from the arena straight away. If an animal did not make a choice within the 2 minute trial time then the animal was removed from the arena and the trial was repeated later.

The position of the positive stimulus was counterbalanced across trials. Animals received 10 trials per session and were considered to have learned the task if they performed at least 75% correct in the last four sessions.

Experiment 2

A subset of animals (those based at University of Lincoln) received a second experiment in which the contingencies associated with the second set of training stimuli were reversed. The procedure was identical to that used in Experiment 1 except that the reward contingencies were reversed, this happened each time the animals reached the learning criterion.

Generalization

After 4 reversals, animals were given a generalization test trained in which they were trained to discriminate between a novel set of stimuli before receiving one reversal with these stimuli. The aim was to assess whether animals were able to generalize what they had learned about reversal to a novel stimulus set. One tortoise did not progress to this phase.

ESM GENERAL METHODS

Repeatability analysis and inclusion criteria for primary data

Repeatability of cognitive performance (R) was computed using the 'rptR' package [1]. This package enables estimation of R for Gaussian, Binomial and Poisson distributed data and additionally provides uncertainty in estimators (95% confidence interval) that are quantified by parametric bootstrapping. It also allowed us to control for fixed effects (e.g. test order, sex, age) and thus assess estimates of adjusted repeatability. Finally, we performed significance testing by conducting likelihood ratio tests.

When necessary, cognitive performance (i.e., accuracy, number of trials to reach criterion, latency to solve a task) was (natural) log or square-root transformed prior to analysis to meet assumptions of normality. When conversion to a Gaussian distribution was not possible, we used a Poisson link function for count data (i.e., number of trials, latency). We used a Binomial link function when cognitive performance was reported as success or failure in a task or trial (i.e. correct or incorrect).

When several measures of cognitive performance were available for the same individuals on the same task, we run an independent repeatability analysis for each measurement and deal with non-independence in the meta-analysis using a covariance matrix that assume 0.5 correlation between those R values [2,3].

We could not normalise or fit an appropriate link function for 5 datasets. In this case, we used a gaussian link function for Generalized Linear Mixed Model (GLMM)-based repeatability computation and only kept R values if models met inclusion criterion described below.

For each possible repeatability analysis, we calculated unadjusted R (i.e. individual as a random intercept but no fixed effects), adjusted R for test order (i.e. individual as random intercept and the repeat number of the test as a fixed effect) and adjusted R for test order and individual determinants (i.e. individual as random intercept and test order and sex and/or age as fixed effects). Some datasets also contained several species, experimental contexts and/or tasks, leading to a total of 208 repeatability analyses.

Each repeatability model was validated by uniformity testing, confirmed by visual inspection of residuals for normality using the DHARMa package [4]. To account for additive overdispersion in Poisson models, we added an observation level random effect [5]. We then ran parametric overdispersion tests using the DHARMa package to verify if overdispersion was still present. R resulting from models that showed non-normal residuals ($p<0.05$, n=27) or overdispersion ($P<0.05$, n=6) were excluded from further analysis (Table S3). We therefore included 178 R values derived from primary datasets in further analysis.

Finally, because repeatabilities based on GLMMs are constrained to be positive, for unadjusted R close to 0 (<0.005), we computed R using the least squares ANOVA approach [3,6] using the 'ICC' package [7].

R was then calculated as: $\frac{\textit{mean squares among repetitions} - \textit{mean squares within repetitions}}{\textit{sample size per group}}$

R values with a negative value occur when there is more variation within individuals than among their means.

Meta -analysis and meta-regression

Moderator descriptions

Cognitive performance measurement was the quantification of a cognitive process using: accuracy, e.g. proportion correct (ACC); the number of trials to reach a learning criterion (TTC); success-or-failure binary outcome (SUC); latency (LAT); normalised performance scores (NOR); the number of correct trials or errors over a fixed number of trials (NBT). Cognitive task type included: 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 include visual exclusion performance; auditory exclusion performance and object permanence; social learning (SL), spatial orientation learning (SOL), spatial recognition (SR) and lexical fluency (LF). Median delay between tests was computed as: the median of delays in days between repetitions of a test for each individual (task beginning used as reference) when information was available in the dataset or was fixed and experimentally defined. Experimental context and the origin of subjects were each comprised of two levels: either

wild (an experiment was conducted in the wild or on wild-caught subjects, respectively) or laboratory (an experiment conducted in a laboratory or on laboratory-raised/hand-raised subjects, respectively). Taxonomic Class is the class (Arachnida, Aves, Gastropoda; Insecta, Mammalia, Reptilia) of each subject. Finally, Publication of R value is a binary variable in which 1 means that this repeatability value has been already published while 0 means the R value has been computed in the present paper from primary data.

Model parameters

For the meta-analyses and meta-regressions, we standardised all repeatability estimates of R using Fisher's Z transformation [8]. Along with the standardised effect size Fisher's ZR $= 0.5ln(\frac{1+(k-1))R}{1-R})$, we calculated the corresponding sampling variances*:* $VarZR = \frac{k}{2(n-1)(k-1)}$. R is the repeatability value, *n* is the total sample size and *k* is the number of repeated tests (individuals performing only 1 test are also taken into account, hence average k could be <2; Table S1 and S2). For all figures (except Figure S10 and S11) and tables, we back-transformed model parameters to their original scale [3], hence ensuring that results remained comparable with a single study on repeatability.

To account for studies that reported multiple cognitive performance measurements for the same experiment, we fitted a variance–covariance matrix, derived from VarZR, in all models to deal with correlation arising from these shared groups. We assumed that the correlations among shared groups were 0.5 [2,3].

Sensitivity analysis

We assessed if results from the meta-analytic model (intercept-only) were driven by a particular study by visually inspecting the distribution of mean effect sizes computed by removing each study one by one (Figure S9).

1. 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.

2. Booksmythe I, Mautz B, Davis J, Nakagawa S, Jennions MD. 2015 Facultative adjustment of the offspring sex ratio and male attractiveness: a systematic review and meta-analysis. *Biol. Rev. Camb. Philos. Soc.* **92**, 108–134.

3. Holtmann B, Lagisz M, Nakagawa S. 2016 Metabolic rates, and not hormone levels, are a likely mediator of between-individual differences in behaviour: a meta-analysis. *Funct. Ecol.* **31**, 685–696.

4. Hartig F. 2017 *DHARMa: Residual Diagnostics for Hierarchical (Multi-Level / Mixed) Regression Models*.

5. Harrison XA. 2014 Using observation-level random effects to model overdispersion in count data in ecology and evolution. *PeerJ* **2**, e616.

6. Lessells CM, Boag PT. 1987 Unrepeatable Repeatabilities: A Common Mistake. *Auk* **104**, 116–121.

7. Wolak ME, Fairbairn DJ, Paulsen YR. 2011 Guidelines for estimating repeatability. *Methods Ecol. Evol.* **3**, 129–137.

8. McGraw KO, Wong SP. 1996 Forming inferences about some intraclass correlation coefficients. *Psychol. Methods* **1**, 30–46.

ESM Results

First Author	Species	Cognitive task	Cognitive performance	Median delay	Experimental condition	Subject Origin	Sample size	Average repetition number	SD repetition number	R	R low CI	R high CI	Rn	Rn lowCI	Rn highCI	Rni	Rni lowCI	Rni highCI
Wilkinson	Chelonoidis carbonaria	RL	TTC	NA	Lab	Lab	4	4.00	0.00	–0.263	–0.3170	0.2580	0.000	0.000	0.301	NA	NA	NA
Chow	Sciurus carolinensis	PS	LAT	660.0	Lab	Lab	5	2.00	0.00	–0.139	–0.8140	0.7530	0.305	0.000	0.570	0.062	0.000	0.860
Van Horik	Pionites melanocephala	RL	TTC	7.0	Lab	Lab	4	7.00	0.00	–0.069	–0.1390	0.4900	0.000	0.000	0.211	NA	NA	NA
Dalesman	Lymnaea stagnalis	ME	NOR	14.0	Lab	Lab	35	3.00	0.00	0.000	0.0000	0.0000	0.044	0.000	0.243	NA	NA	NA
Barbeau	Homo sapiens	ME	NBT	540.0	Lab	Wild	41	1.78	0.42	0.000	0.0000	0.1120	0.000	0.000	0.097	NA	NA	NA
Cauchoix	Parus major	RL	ACC	0.0	Wild	Wild	20	27.25	17.86	0.012	0.0000	0.0570	0.000	0.000	0.000	NA	NA	NA
Cauchoix	Parus major	RL	ACC	0.0	Lab	Wild	20	16.65	5.66	0.016	0.0000	0.0420	0.015	0.000	0.049	0.002	0.000	0.065
Cauchoix	Parus major	RL	TTC	0.0	Lab	Wild	20	16.65	5.66	0.039	0.0010	0.1020	0.036	0.000	0.073	0.035	0.000	0.097
Matzel	Mus musculus	LE	NOR	7.0	Lab	Lab	56	1.14	0.35	0.074	0.0000	0.6280	0.168	0.000	0.721	NA	NA	NA
Matzel	Mus musculus	DL	NOR	7.0	Lab	Lab	56	1.14	0.35	0.098	0.0000	0.6380	0.623	0.056	0.877	NA	NA	NA
Van Horik	Diopsittaca nobilis	RL	TTC	7.0	Lab	Lab	4	7.00	0.00	0.104	0.0000	0.5060	0.096	0.000	0.409	0.098	0.000	0.232
Schuster	Micromys minutus	SOL	PRO	139.0	Lab	Lab	56	2.00	NA	0.125	0.0000	0.4110	NA	NA	NA	NA	NA	NA
Cauchard	Parus major	PS	LAT	360.0	Wild	Wild	209	1.07	0.30	0.163	0.0000	0.5680	0.092	0.000	0.438	0.000	0.000	0.421
Schuster	Micromys minutus	SR	LAT	76.1	Lab	Lab	96	2.00	NA	0.200	0.0000	0.4290	NA	NA	NA	NA	NA	NA
Schuster	Micromys minutus	SR	LAT	84.0	Lab	Lab	31	2.00	NA	0.200	0.0000	0.5300	NA	NA	NA	NA	NA	NA
Matzel	Mus musculus	SOL	NOR	7.0	Lab	Lab	56	1.14	0.35	0.207	0.0000	0.6830	0.192	0.000	0.740	NA	NA	NA
Claidiere	Papio papio	ME	ACC	70.0	Lab	Lab	20	2.20	0.89	0.219	0.0800	0.3960	0.681	0.413	0.823	0.724	0.579	0.854
Nawroth	Capra hircus	PC	ACC	14.0	Lab	Lab	11	2.00	0.00	0.220	0.0000	0.5230	0.524	0.189	0.844	0.555	0.322	0.813
Cole	Parus major	PS	SUC	365.0	Lab	Wild	35	2.00	NA	0.270	–0.0436	0.5836	NA	NA	NA	NA	NA	NA
Bize	Ficedula albicollis	PS	LAT	1.0	Wild	Wild	375	1.28	0.45	0.272	0.1290	0.3270	0.318	0.202	0.405	0.317	0.249	0.425
Matzel	Mus musculus	LE	NOR	7.0	Lab	Lab	56	1.14	0.35	0.318	0.0000	0.5370	0.715	0.670	0.866	NA	NA	NA
Cole	Parus major	PS	SUC	1.0	Lab	Wild	347	2.00	NA	0.340	0.2420	0.4380	NA	NA	NA	NA	NA	NA
Cole	Parus major	PS	SUC	1.0	Lab	Wild	80	2.00	NA	0.370	0.1740	0.5660	NA	NA	NA	NA	NA	NA
Cole	Parus major	PS	SUC	1095.0	Lab	Wild	47	2.00	NA	0.370	0.1152	0.6248	NA	NA	NA	NA	NA	NA
Cole	Parus major	PS	SUC	1095.0	Lab	Wild	67	2.00	NA	0.400	0.2040	0.5960	NA	NA	NA	NA	NA	NA
Barbeau	Homo sapiens	ME	ACC	540.0	Lab	Wild	40	1.60	0.50	0.420	0.0610	0.6770	0.409	0.364	0.747	0.311	0.104	0.593
Barbeau	Homo sapiens	ME	NOR	540.0	Lab	Wild	41	1.66	0.48	0.421	0.2370	0.5990	0.566	0.410	0.696	0.563	0.204	0.670
Barbeau	Homo sapiens	ME	NBT	540.0	Lab	Wild	41	1.76	0.43	0.498	0.3100	0.5910	0.483	0.362	0.640	0.475	0.182	0.695
Barbeau	Homo sapiens	LF	NBT	540.0	Lab	Wild	41	1.73	0.45	0.505	0.0280	0.7740	0.476	0.000	0.642	NA	NA	NA
Langbein	Capra hircus	RL	TTC	14.0	Lab	Lab	75	2.56	0.72	0.519	0.4420	0.6120	0.641	0.520	0.726	NA	NA	NA
Cole	Parus major	PS	SUC	365.0	Lab	Wild	23	2.00	NA	0.540	0.2460	0.8340	NA	NA	NA	NA	NA	NA
Barbeau	Homo sapiens	ME	NBT	540.0	Lab	Wild	41	1.78	0.42	0.561	0.5110	0.7130	0.552	0.329	0.642	0.500	0.426	0.650
Cauchard	Parus major	PS	ACC	360.0	Wild	Wild	350	1.09	0.30	0.598	0.4850	0.7550	0.564	0.428	0.696	0.556	0.341	0.696
Barbeau	Homo sapiens	ME	NBT	540.0	Lab	Wild	41	1.76	0.43	0.650	0.4600	0.7700	0.639	0.334	0.730	0.608	0.475	0.710
Barbeau	Homo sapiens	LF	NBT	540.0	Lab	Wild	41	1.73	0.45	0.657	0.3020	0.8210	0.693	0.637	0.807	0.700	0.455	0.814
Barbeau	Homo sapiens	ME	NOR	540.0	Lab	Wild	41	1.68	0.47	0.684	0.5360	0.8530	0.677	0.488	0.798	0.660	0.635	0.817
Barbeau	Homo sapiens	ME	NBT	540.0	Lab	Wild	41	1.66	0.48	0.696	0.5670	0.8390	0.703	0.578	0.836	0.695	0.510	0.852
Barbeau	Homo sapiens	ME	NBT	540.0	Lab	Wild	40	1.80	0.41	0.735	0.5940	0.8350	0.751	0.637	0.883	0.713	0.562	0.878
Matzel	Mus musculus	SOL	NOR	7.0	Lab	Lab	56	1.14	0.35	0.752	0.4280	0.9040	0.740	0.641	0.910	NA	NA	NA
Barbeau	Homo sapiens	ME	NBT	540.0	Lab	Wild	40	1.80	0.41	0.812	0.6620	0.8390	0.823	0.774	0.841	0.776	0.671	0.886
Cauchoix	Parus major	RL	TTC	0.0	Wild	Wild	20	27.25	17.86	NA	0.0690	0.1600	0.075	0.049	0.129	NA	NA	NA
Cauchoix	Parus major	RL	ACC	0.0	Lab	Wild	17	30.76	26.13	NA	0.0530	0.2150	0.020	0.000	0.058	NA	NA	NA
Cauchoix	Parus major	RL	TTC	0.0	Lab	Wild	17	30.76	26.13	NA	0.0850	0.3560	NA	0.067	0.203	NA	NA	NA
Nawroth	Sus scrofa	HC	NBT	7.0	Lab	Lab	17	1.76	0.44	NA	0.0000	0.1390	NA	0.000	0.069	NA	0.000	0.139
Lihoreau	Bombus terrestris	SOL	NOR	0.0	Wild	Lab	7	29.00	5.80	NA	0.0000	0.0520	0.076	0.012	0.231	NA	NA	NA
Lihoreau	Bombus terrestris	SOL	NOR	0.0	Lab	Lab	7	79.86	0.38	NA	0.0070	0.0420	0.018	0.005	0.058	NA	NA	NA
Lihoreau	Bombus terrestris	SOL	NOR	0.0	Lab	Lab	10	40.00	0.00	NA	0.1140	0.3310	0.184	0.077	0.342	NA	NA	NA
Dalesman	Lymnaea stagnalis	ME	NOR	14.0	Lab	Lab	40	2.00	0.00	NA	0.2720	0.6610	0.557	0.297	0.737	NA	NA	NA
Huebner	Microcebus murinus	PS	LAT	341.0	Lab	Wild	82	1.27	0.56	NA	0.3080	0.7520	NA	0.486	0.751	NA	0.498	0.868
Huebner	Microcebus murinus	PS	LAT	341.0	Lab	Wild	83	1.24	0.53	NA	0.0000	0.3450	NA	0.000	0.275	0.000	0.000	0.393
Barbeau	Homo sapiens	ME	NBT	540.0	Lab	Wild	41	1.78	0.42	NA	0.0000	0.0300	NA	0.000	0.149	NA	NA	NA
Barbeau	Homo sapiens	ME	NBT	540.0	Lab	Wild	41	1.76	0.43	NA	0.2570	0.5700	0.545	0.278	0.653	NA	NA	NA
Schuster	Micromys minutus	SOL	TTC	139.0	Lab	Lab	57	2.00	NA	NA	NA	NA	0.127	0.000	0.382	NA	NA	NA
Schuster	Micromys minutus	SOL	LAT	139.0	Lab	Lab	56	2.00	NA	NA	NA	NA	0.263	0.211	0.313	NA	NA	NA
Schuster	Micromys minutus	SR	LAT	7.0	Lab	Lab	39	2.00	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Rodriguez	Frontinella communis	ME	LAT	3.0	Lab	Wild	33	2.00	NA	NA	NA	NA	NA	NA	NA	0.420	NA	NA
Rodriguez	Frontinella communis	ME	LAT	3.0	Lab	Wild	33	2.00	NA	NA	NA	NA	NA	NA	NA	0.600	NA	NA
Rodriguez	Frontinella communis	ME	LAT	3.0	Lab	Wild	33	2.00	NA	NA	NA	NA	NA	NA	NA	0.520	NA	NA
Rodriguez	Frontinella communis	ME	LAT	3.0	Lab	Wild	33	2.00	NA	NA	NA	NA	NA	NA	NA	0.160	NA	NA
Rodriguez	Frontinella communis	ME	LAT	3.0	Lab	Wild	33	2.00	NA	NA	NA	NA	NA	NA	NA	–0.170	NA	NA
Rodriguez	Frontinella communis	ME	LAT	3.0	Lab	Wild	33	2.00	NA	NA	NA	NA	NA	NA	NA	–0.070	NA	NA

Table S1: Temporal repeatability. Data are ordered by ascending R values.

R = unadjusted repeatability; Rn = repeatability adjusted for test order; Rni = repeatability adjusted for test order and individual determinants. Cognitive performance measurement was the quantification of a cognitive process using: accuracy, e.g. proportion correct (ACC); the number of trials to reach a learning criterion (TTC); success-or-failure binary outcome (SUC); latency (LAT); normalised performance scores (NOR); the number of correct trials or errors over a fixed number of trials (NBT). Cognitive task type included: 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 include visual exclusion performance; auditory exclusion performance and object permanence; social learning (SL), spatial orientation learning (SOL), spatial recognition (SR) and lexical fluency (LF). Median delay between tests was computed as: the median of delays in days between repetitions of a test for each individual (task beginning used as reference) when information was available in the dataset or was fixed and experimentally defined. Experimental context and the origin of subjects were each comprised of two levels: either wild (an experiment was conducted in the wild or on wild-caught subjects, respectively) or laboratory (an experiment conducted in a laboratory or on laboratory-raised/hand-raised subjects, respectively). We replaced Rs values by NA when the model was not fulfilling inclusion criterion (see methodes) or when no information about individual was available for Rni column only.

First Author	Species	Cognitive task	Cognitive performance	Median delay	Experimental condition	Subject Origin	Sample size	Average repetition number	SD repetition number	R	R low CI	R high CI	Rn	Rn lowCI	Rn highCI	Rni	Rni lowCI	Rni highCI
Wilkinson	Chelonoidis carbonaria	RL_RL	TTC	NA	Lab	Lab	4	1.75	0.50	−0.586	−1.323	0.731	0.905	0.769	0.998	NA	NA	NA
Henke−v.d.Malsburg	Microcebus murinus	RL_RL	ACC	1	Lab	Wild	6	2.00	0.00	−0.497	−0.894	0.402	0.000	0.000	0.406	0.000	0.000	0.545
Nawroth	Capra hircus	PC_PC	ACC	390	Lab	Lab	11	1.55	0.52	−0.328	−1.377	0.501	0.064	0.000	0.782	0.000	0.000	0.000
Henke−v.d.Malsburg	Microcebus berthae	PS_PS	LAT	1	Lab	Wild	11	2.55	0.69	−0.302	−0.514	0.147	0.000	0.000	0.127	0.000	0.000	0.221
Henke−v.d.Malsburg	Microcebus murinus	DL_DL	TTC	1	Lab	Wild	7	1.86	0.38	−0.223	−0.923	0.608	0.000	0.000	0.444	0.000	0.000	0.577
Chow	Sciurus vulgaris	PS_PS	LAT	5	Wild	Wild	14	2.00	0.00	−0.194	−0.634	0.351	0.000	0.000	0.567	0.000	0.000	0.377
Henke−v.d.Malsburg	Microcebus murinus	DL_DL	ACC	1	Lab	Wild	7	1.86	0.38	−0.176	−0.902	0.635	0.000	0.000	0.647	0.031	0.000	0.853
Nawroth	Capra hircus	PC_PC	ACC	7	Lab	Lab	10	3.00	0.00	−0.121	−0.340	0.330	0.313	0.000	0.691	0.327	0.040	0.564
Henke−v.d.Malsburg	Microcebus murinus	PS_PS	LAT	1	Lab	Wild	20	2.30	0.80	−0.118	−0.414	0.269	0.143	0.000	0.373	0.168	0.005	0.461
Chow	Sciurus carolinensis	PS_PS	TTC	6	Lab	Lab	5	2.00	0.00	−0.017	−0.768	0.801	0.541	0.054	0.978	0.512	0.341	0.943
Chow	Sciurus carolinensis	DL_DL	TTC	466	Lab	Lab	5	2.60	0.55	0.000	0.000	0.464	0.757	0.225	0.840	0.522	0.000	0.791
Huebner	Microcebus murinus	PS_PS	LAT	0	Lab	Wild	91	1.77	0.42	0.000	0.000	0.237	0.000	0.000	0.201	0.000	0.000	0.135
Shaw	Petroica longipes	IN_IN	TTC	365	Wild	Wild	15	2.00	NA	0.002	0.000	0.267	NA	NA	NA	NA	NA	NA
Shaw	Petroica longipes	LE_LE	TTC	365	Wild	Wild	16	2.00	NA	0.020	0.000	0.478	NA	NA	NA	NA	NA	NA
Henke−v.d.Malsburg	Microcebus murinus	PS_PS	ACC	1	Lab	Wild	20	2.30	0.80	0.024	−0.309	0.406	0.000	0.000	0.218	0.000	0.000	0.196
Van Horik	Phasianus colchicus	DL_DL	NBT	5	Lab	Lab	158	2.00	0.00	0.040	0.000	0.202	0.037	0.000	0.079	0.034	0.000	0.193
Chow	Sciurus carolinensis	PS_PS	LAT	5	Wild	Wild	11	2.00	0.00	0.112	0.000	0.533	0.068	0.000	0.494	0.094	0.000	0.481
Barragan−Jason	Homo sapiens	IN_IN	SUC	7	Wild	Wild	43	1.58	0.50	0.132	0.003	0.974	0.142	0.002	0.658	0.117	0.000	0.865
Nawroth	Capra hircus	HC_HC	ACC	390	Lab	Lab	11	1.91	0.30	0.140	0.000	0.643	0.119	0.000	0.559	0.090	0.000	0.617
Cabirol	Apis mellifera	DL_DL	SUC	0	Lab	Wild	47	2.00	0.00	0.147	0.000	0.372	0.147	0.000	0.232	NA	NA	NA
Cauchoix	Parus major	RL_RL	TTC	7	Lab	Wild	20	1.45	0.51	0.198	0.000	0.747	0.544	0.140	0.880	0.541	0.009	0.850
Nawroth	Sus scrofa	HC_HC	NBT	7	Lab	Lab	17	4.00	0.00	0.205	0.075	0.362	0.217	0.007	0.468	0.228	0.013	0.399
Matzel	Mus musculus	IA_EA	NOR	22	Lab	Lab	26	4.00	0.00	0.213	0.020	0.289	0.208	0.035	0.273	NA	NA	NA
Cole	Parus major	PS_PS	SUC	1	Lab	Wild	297	2.00	NA	0.240	0.142	0.338	NA	NA	NA	NA	NA	NA
Matzel	Mus musculus	IA_IA	NOR	22	Lab	Lab	26	2.00	0.00	0.253	0.113	0.554	0.235	0.000	0.475	NA	NA	NA
Langley	Phasianus colchicus	DL_DL	ACC	6	Lab	Lab	39	2.00	0.00	0.261	0.041	0.483	0.254	0.066	0.471	0.259	0.123	0.391
Cole	Parus major	PS_PS	SUC	365	Lab	Wild	23	2.00	NA	0.310	0.212	0.408	NA	NA	NA	NA	NA	NA
Cauchoix	Parus major	RL_RL	ACC	7	Lab	Wild	20	1.45	0.51	0.343	0.095	0.561	0.000	0.000	0.608	0.000	0.000	0.871
Cole	Parus major	PS_PS	SUC	1095	Lab	Wild	46	2.00	NA	0.350	0.252	0.448	NA	NA	NA	NA	NA	NA
Klein	Bombus terrestris	SOL_SOL	NOR	1	Lab	Lab	29	3.00	0.00	0.361	0.176	0.533	0.480	0.311	0.581	NA	NA	NA
Sorato	Gallus gallus	DL_DL	TTC	180	Lab	Lab	113	1.25	0.43	0.465	0.209	0.735	0.471	0.281	0.557	0.438	0.117	0.674
Matzel	Mus musculus	EA_EA	NOR	22	Lab	Lab	26	2.00	0.00	0.577	0.370	0.746	0.564	0.350	0.705	NA	NA	NA
Wilkinson	Chelonoidis carbonaria	DL_DL	TTC	NA	Lab	Lab	7	1.86	0.38	0.580	0.056	0.800	0.593	0.118	0.880	NA	NA	NA
Henke−v.d.Malsburg	Microcebus berthae	PS_PS	ACC	1	Lab	Wild	11	2.55	0.69	0.587	0.414	0.765	0.614	0.268	0.817	0.639	0.352	0.810
Ashton	Cracticus tibicen dorsalis	IN_IN	NBT	14	Wild	Wild	56	2.00	NA	0.806	0.691	0.882	NA	NA	NA	NA	NA	NA
Ashton	Cracticus tibicen dorsalis	ME_ME	NBT	14	Wild	Wild	46	2.00	NA	0.932	0.879	0.963	NA	NA	NA	NA	NA	NA
Ashton	Cracticus tibicen dorsalis	DL_DL	TTC	14	Wild	Wild	46	2.00	NA	0.970	0.946	0.983	NA	NA	NA	NA	NA	NA
Ashton	Cracticus tibicen dorsalis	RL_RL	TTC	14	Wild	Wild	46	2.00	NA	0.975	0.954	0.986	NA	NA	NA	NA	NA	NA
Dalesman	Lymnaea stagnalis	ME_ME	NOR	7	Lab	Lab	80	2.00	0.00	NA	0.133	0.373	0.290	0.092	0.503	NA	NA	NA
Chow	Sciurus carolinensis	RL_RL	TTC	464	Lab	Lab	5	2.60	0.55	NA	0.000	0.238	0.142	0.000	0.698	0.000	0.000	0.270
Huber	Canis lupus familiaris	DL_DL	TTC	30	Lab	Wild	105	2.71	0.63	NA	0.000	0.028	NA	0.068	0.277	0.089	0.037	0.225
Huber	Canis lupus familiaris	DL_DL	NBT	30	Lab	Wild	105	2.71	0.63	NA	0.000	0.066	0.059	0.000	0.143	0.014	0.000	0.092
Atance	Homo sapiens	ME_ME	NBT	0	Lab	Wild	92	1.96	0.33	NA	0.203	0.562	NA	0.315	0.533	0.016	0.000	0.154
Atance	Homo sapiens	IN_IN	NBT	0	Lab	Wild	92	2.01	0.23	NA	−0.256	0.145	NA	0.000	0.000	0.000	0.000	0.198
Hanson	Homo sapiens	ME_ME	NBT	0	Lab	Wild	86	1.97	0.18	NA	0.497	0.665	NA	0.407	0.606	0.272	0.166	0.393
Guenther	Cavia aperea	PS_PS	LAT	12	Lab	Lab	24	4.00	NA	NA	NA	NA	NA	NA	NA	0.500	0.330	0.780
Guenther	Cavia aperea	PS_PS	SUC	12	Lab	Lab	24	4.00	NA	NA	NA	NA	NA	NA	NA	0.450	0.180	0.700
Guenther	Cavia aperea	DL_DL	TTC	40	Lab	Lab	24	4.00	NA	NA	NA	NA	NA	NA	NA	0.060	0.000	0.320
Guenther	Cavia aperea	DL_DL	SUC	40	Lab	Lab	24	4.00	NA	NA	NA	NA	NA	NA	NA	0.860	0.400	0.870
Guenther	Cavia aperea	SL_SL	SUC	50	Lab	Lab	24	4.00	NA	NA	NA	NA	NA	NA	NA	0.300	0.150	0.380
Guenther	Cavia aperea	SL_SL	NOR	50	Lab	Lab	24	4.00	NA	NA	NA	NA	NA	NA	NA	0.450	0.250	0.670
Rodriguez	Frontinella communis	ME_ME	LAT	3	Lab	Wild	33	4.00	NA	NA	NA	NA	NA	NA	NA	0.250	NA	NA
Rodriguez	Frontinella communis	ME_ME	LAT	3	Lab	Wild	33	4.00	NA	NA	NA	NA	NA	NA	NA	0.170	NA	NA
Rodr1guez	Frontinella communis	ME_ME	LAT	3	Lab	Wild	33	4.00	NA	NA	NA	NA	NA	NA	NA	0.170	NA	NA

Table S2: Contextual repeatability. Data are ordered by ascending unadjusted R values.
R = unadjusted repeatability; Rn = repeatability adjusted for test order; Rni = repeatability adjusted for test order and individual determinants. Cognitive performance measurement was the quantification of a cognitive process using: accuracy, e.g. proportion correct (ACC); the number of trials to reach a learning criterion (TTC); success-or-failure binary outcome (SUC); latency (LAT); normalised performance scores (NOR); the number of correct trials or errors over a fixed number of trials (NBT). Cognitive task type included: 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 include visual exclusion performance; auditory exclusion performance and object permanence; social learning (SL), spatial orientation learning (SOL), spatial recognition (SR) and lexical fluency (LF). Median delay between tests was computed as: the median of delays in days between repetitions of a test for each individual (task beginning used as reference) when information was available in the dataset or was fixed and experimentally defined. Experimental context and the origin of subjects were each comprised of two levels: either wild (an experiment was conducted in the wild or on wild-caught subjects, respectively) or laboratory (an experiment conducted in a laboratory or on laboratory-raised/hand-raised subjects, respectively). We replaced Rs values by NA when the model was not fulfilling inclusion criterion (see methodes) or when no information about individual was available for Rni column only.

First Author	Laboratory	Analysis Name	Species	Cognitive performance	Exclusion	R type	R Published	Publication category	Link function	R mod conv	R unif (pval)	R over (pval)	Rn mod conv	Rn unif (pval)	Rn over (pval)	Rni mod conv	Rni unif (pval)	Rni over (pval)
Ashton	Thornton	Ashton_IN	Cracticus tibicen dorsalis	NBT	NA	Contextual	yes	forR	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Ashton	Thornton	Ashton_DL	Cracticus tibicen dorsalis	TTC	NA	Contextual	yes	forR	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Ashton	Thornton	Ashton_RL	Cracticus tibicen dorsalis	TTC	NA	Contextual	yes	forR	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Ashton	Thornton	Ashton_ME	Cracticus tibicen dorsalis	NBT	NA	Contextual	yes	forR	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Atance	Atance	Atance_2017_children_WM	Homo sapiens	NBT	R Rn	Contextual	no	notForR	Gaussian	1	0.000	NA	1	0.000	NA	1	0.715	NA
Atance	Atance	Atance_2017_children_Inhibition	Homo sapiens	NBT	R Rn	Contextual	no	notForR	Gaussian	1	0.000	NA	1	0.000	NA	1	0.000	NA
Barbeau	Barbeau	Barbeau_Rec.Visage_MULTI	Homo sapiens	NOR	No	Temporal	no	notForR	Gaussian	1	0.118	NA	1	0.075	NA	1	0.514	NA
Barbeau	Barbeau	Barbeau_Rec.Visage_MULTI	Homo sapiens	NOR	No	Temporal	no	notForR	Gaussian	1	0.394	NA	1	0.336	NA	1	0.224	NA
Barbeau	Barbeau	Barbeau_Rec.verb..score	Homo sapiens	ACC	No	Temporal	no	notForR	Gaussian	1	0.706	NA	1	0.834	NA	1	0.743	NA
Barbeau	Barbeau	Barbeau_Evoc.lex_MULTI	Homo sapiens	NBT	No	Temporal	no	notForR	Gaussian	1	0.891	NA	1	0.464	NA	1	0.349	NA
Barbeau	Barbeau	Barbeau_G.B_MULTI	Homo sapiens	NBT	No	Temporal	no	notForR	Gaussian	1	0.835	NA	1	0.942	NA	1	0.840	NA
Barbeau	Barbeau	Barbeau_G.B_MULTI	Homo sapiens	NBT	No	Temporal	no	notForR	Gaussian	1	0.992	NA	1	1.000	NA	1	0.991	NA
Barbeau	Barbeau	Barbeau_G.B_MULTI	Homo sapiens	NBT	No	Temporal	no	notForR	Gaussian	1	0.157	NA	1	0.318	NA	1	0.760	NA
Barbeau	Barbeau	Barbeau_Short.Eve_MULTI	Homo sapiens	NBT	No	Temporal	no	notForR	Gaussian	1	0.374	NA	1	0.553	NA	1	0.821	NA
Barbeau	Barbeau	Barbeau_Vis.c_l.1	Homo sapiens	NBT	No	Temporal	no	notForR	Gaussian	1	0.811	NA	1	0.900	NA	1	0.783	NA
Barbeau	Barbeau	Barbeau_Short.Eve_MULTI	Homo sapiens	NBT	No	Temporal	no	notForR	Gaussian	1	0.318	NA	1	0.587	NA	1	0.169	NA
Barbeau	Barbeau	Barbeau_Empan_MULTI	Homo sapiens	NBT	R Rn	Temporal	no	notForR	Poisson	1	0.009	1.000	1	0.007	1.000	NA	NA	NA
Barbeau	Barbeau	Barbeau_Empan_MULTI	Homo sapiens	NBT	No	Temporal	no	notForR	Poisson	1	0.089	0.982	1	0.137	0.986	NA	NA	NA
Barbeau	Barbeau	Barbeau_Evoc.lex_MULTI	Homo sapiens	NBT	No	Temporal	no	notForR	Poisson	1	0.969	0.799	1	0.269	0.599	NA	NA	NA
Barbeau	Barbeau	Barbeau_G.B_MULTI	Homo sapiens	NBT	R	Temporal	no	notForR	Poisson	1	0.769	0.048	1	0.183	0.506	NA	NA	NA
Barragan–Jason	Barragan	Barragan–Jason_b_1st	Homo sapiens	SUC	No	Contextual	no	Methods	Binary	1	0.671	NA	1	0.502	NA	1	0.403	NA
Bize	Doligez	Bize_flycatcher_ps	Ficedula albicollis	LAT	No	Temporal	no	Methods	Gaussian	1	0.096	NA	1	0.145	NA	1	0.198	NA
Cabirol	Devaud	Cabirol_differential_finalScoreS+	Apis mellifera	SUC	No	Contextual	no	notForR	Binary	1	0.924	NA	1	0.933	NA	NA	NA	NA
Cauchard	Doligez	Cauchard_tits_ps_MULTI	Parus major	ACC	No	Temporal	no	notForR	Gaussian	1	0.875	NA	1	0.781	NA	1	0.488	NA
Cauchard	Doligez	Cauchard_tits_ps_MULTI	Parus major	LAT	No	Temporal	no	notForR	Poisson	1	0.781	0.406	1	0.960	0.433	1	0.978	0.099
Cauchoix	Chaine	Cauchoix_Single_MULTI	Parus major	ACC	No	Temporal	no	Methods	Gaussian	1	0.993	NA	1	0.998	NA	1	0.950	NA
Cauchoix	Chaine	Cauchoix_Single_MULTI	Parus major	TTC	No	Temporal	no	Methods	Gaussian	1	0.822	NA	1	0.973	NA	1	0.984	NA
Cauchoix	Chaine	Cauchoix_wild_MULTI	Parus major	ACC	No	Temporal	no	notForR	Gaussian	1	0.892	NA	1	0.931	NA	NA	NA	NA
Cauchoix	Chaine	Cauchoix_wild_MULTI	Parus major	TTC	R	Temporal	no	notForR	Gaussian	1	0.011	NA	1	0.318	NA	NA	NA	NA
Cauchoix	Chaine	Cauchoix_captivity_MULTI	Parus major	ACC	R	Temporal	no	notForR	Gaussian	1	0.009	NA	1	0.955	NA	NA	NA	NA
Cauchoix	Chaine	Cauchoix_captivity_MULTI	Parus major	TTC	R Rn	Temporal	no	notForR	Gaussian	1	0.000	NA	1	0.005	NA	NA	NA	NA
Cauchoix	Chaine	Cauchoix_Single_spatial2color_MULTI	Parus major	ACC	No	Contextual	no	Methods	Gaussian	1	0.738	NA	1	0.590	NA	1	0.671	NA
Cauchoix	Chaine	Cauchoix_Single_spatial2color_MULTI	Parus major	TTC	No	Contextual	no	Methods	Gaussian	1	0.821	NA	1	0.579	NA	1	0.527	NA
Chow	Lea	Chow_lab_samePS	Sciurus carolinensis	LAT	No	Temporal	no	notForR	Gaussian	1	0.210	NA	1	0.643	NA	1	0.869	NA
Chow	Lea	Chow_gray_squirrel_wild	Sciurus carolinensis	LAT	No	Contextual	no	notForR	Gaussian	1	0.956	NA	1	0.883	NA	1	0.987	NA
Chow	Lea	Chow_red_squirrel_wild_latency	Sciurus vulgaris	LAT	No	Contextual	no	notForR	Gaussian	1	0.815	NA	1	0.863	NA	1	0.739	NA
Chow	Lea	Chow_lab_DL	Sciurus carolinensis	TTC	No	Contextual	no	notForR	Poisson	1	0.556	0.117	1	0.604	0.052	1	0.671	0.057
Chow	Lea	Chow_lab_RL	Sciurus carolinensis	TTC	R	Contextual	no	notForR	Poisson	1	0.586	0.047	1	0.773	0.082	1	0.672	0.449
Chow	Lea	Chow_lab_diffPS	Sciurus carolinensis	TTC	No	Contextual	no	notForR	Gaussian	1	0.454	NA	1	0.882	NA	1	0.780	NA
Claidiere	Fagot	Claidiere_summary	Papio papio	ACC	No	Temporal	no	notForR	Gaussian	1	0.542	NA	1	0.193	NA	1	0.055	NA
Cole	Quinn	Cole_withinCapt_LP	Parus major	SUC	NA	Temporal	yes	forR	Binary	NA	NA	NA	NA	NA	NA	NA	NA	NA
Cole	Quinn	Cole_withinCapt_SP	Parus major	SUC	NA	Temporal	yes	forR	Binary	NA	NA	NA	NA	NA	NA	NA	NA	NA
Cole	Quinn	Cole_betweenCapt_1year_LP	Parus major	SUC	NA	Temporal	yes	forR	Binary	NA	NA	NA	NA	NA	NA	NA	NA	NA
Cole	Quinn	Cole_betweenCapt_3year_LP	Parus major	SUC	NA	Temporal	yes	forR	Binary	NA	NA	NA	NA	NA	NA	NA	NA	NA
Cole	Quinn	Cole_betweenCapt_1year_SP	Parus major	SUC	NA	Temporal	yes	forR	Binary	NA	NA	NA	NA	NA	NA	NA	NA	NA
Cole	Quinn	Cole_betweenCapt_3year_SP	Parus major	SUC	NA	Temporal	yes	forR	Binary	NA	NA	NA	NA	NA	NA	NA	NA	NA
Cole	Quinn	Cole_withinCapt	Parus major	SUC	NA	Contextual	yes	forR	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Cole	Quinn	Cole_betweenCapt_1year	Parus major	SUC	NA	Contextual	yes	forR	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Cole	Quinn	Cole_betweenCapt_3year	Parus major	SUC	NA	Contextual	yes	forR	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Dalesman	Dalesman	Dalesman_opp	Lymnaea stagnalis	NOR	No	Temporal	no	unpub	Poisson	1	0.081	0.611	1	0.395	0.541	NA	NA	NA
Dalesman	Dalesman	Dalesman_app	Lymnaea stagnalis	NOR	R	Temporal	no	notForR	Poisson	1	0.031	0.057	1	0.081	0.197	NA	NA	NA
Dalesman	Dalesman	Dalesman_app_ave	Lymnaea stagnalis	NOR	R	Contextual	no	notForR	Gaussian	1	0.031	NA	1	0.145	NA	NA	NA	NA
Guenther	Brust	Guenther_PS	Cavia aperea	LAT	NA	Contextual	yes	forR	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Guenther	Brust	Guenther_PS	Cavia aperea	SUC	NA	Contextual	yes	forR	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Guenther	Brust	Guenther_DL	Cavia aperea	TTC	NA	Contextual	yes	forR	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Guenther	Brust	Guenther_DL	Cavia aperea	SUC	NA	Contextual	yes	forR	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Guenther	Brust	Guenther_SL	Cavia aperea	NOR	NA	Contextual	yes	forR	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Hanson	Atance	Hanson_children_WM	Homo sapiens	NBT	R Rn	Contextual	no	notForR	Gaussian	1	0.000	NA	1	0.000	NA	1	0.542	NA
Henke–v.d.Malsburg	Fichtel	Henke–v.d.Malsburg_discrimination_MULTI	Microcebus murinus	ACC	No	Contextual	no	unpub	Gaussian	1	0.290	NA	1	0.704	NA	1	0.825	NA
Henke–v.d.Malsburg	Fichtel	Henke–v.d.Malsburg_discrimination_MULTI	Microcebus murinus	TTC	No	Contextual	no	unpub	Gaussian	1	0.744	NA	1	0.801	NA	1	0.974	NA
Henke–v.d.Malsburg	Fichtel	Henke–v.d.Malsburg_reversal	Microcebus murinus	ACC	No	Contextual	no	unpub	Gaussian	1	0.775	NA	1	0.933	NA	1	0.968	NA
Henke–v.d.Malsburg	Fichtel	Henke–v.d.Malsburg_murinus_MULTI	Microcebus murinus	ACC	No	Contextual	no	unpub	Gaussian	1	0.233	NA	1	0.252	NA	1	0.475	NA
Henke–v.d.Malsburg	Fichtel	Henke–v.d.Malsburg_murinus_MULTI	Microcebus murinus	LAT	No	Contextual	no	unpub	Gaussian	1	0.717	NA	1	0.823	NA	1	0.805	NA
Henke–v.d.Malsburg	Fichtel	Henke–v.d.Malsburg_berthae_MULTI	Microcebus berthae	ACC	No	Contextual	no	unpub	Gaussian	1	0.608	NA	1	0.265	NA	1	0.489	NA
Henke–v.d.Malsburg	Fichtel	Henke–v.d.Malsburg_berthae_MULTI	Microcebus berthae	LAT	No	Contextual	no	unpub	Gaussian	1	0.293	NA	1	0.537	NA	1	0.458	NA
Huber	Huber	Huber_dog_MULTI	Canis lupus familiaris	TTC	R Rn	Contextual	no	notForR	Poisson	1	0.025	0.085	1	0.037	0.985	1	0.114	0.990
Huber	Huber	Huber_dog_MULTI	Canis lupus familiaris	NBT	R	Contextual	no	notForR	Poisson	1	0.006	0.864	1	0.338	0.345	1	0.246	0.646
Huebner	Kappeler	Huebner_FE_Success	Microcebus murinus	LAT	R Rn Rni	Temporal	no	notForR	Poisson	1	0.004	0.011	1	0.023	0.019	1	0.013	0.072
Huebner	Kappeler	Huebner_SP_Success	Microcebus murinus	LAT	R Rn	Temporal	no	notForR	Poisson	1	0.425	0.009	1	0.201	0.013	1	0.139	0.097
Huebner	Kappeler	Huebner_SP_FE_Success	Microcebus murinus	LAT	No	Contextual	no	notForR	Poisson	1	0.181	0.087	1	0.659	0.397	1	0.656	0.274
Klein	Lihoreau	Klein_spatialConfig	Bombus terrestris	NOR	No	Contextual	no	notForR	Gaussian	1	0.261	NA	1	0.199	NA	NA	NA	NA
Langbein	Langbein	Langbein_goat_all	Capra hircus	TTC	No	Temporal	no	unpub	Gaussian	1	0.229	NA	1	0.083	NA	NA	NA	NA
Langley	Madden	Langley_pheasant	Phasianus colchicus	ACC	No	Contextual	no	unpub	Gaussian	1	0.707	NA	1	0.857	NA	1	0.694	NA
Lihoreau	Lihoreau	Lihoreau_plosB	Bombus terrestris	NOR	R	Temporal	no	notForR	Gaussian	1	0.032	NA	1	0.613	NA	NA	NA	NA
Lihoreau	Lihoreau	Lihoreau_bioLett	Bombus terrestris	NOR	R	Temporal	no	notForR	Gaussian	1	0.012	NA	1	0.609	NA	NA	NA	NA
Lihoreau	Lihoreau	Lihoreau_functEcol	Bombus terrestris	NOR	R	Temporal	no	notForR	Gaussian	1	0.000	NA	1	0.421	NA	NA	NA	NA
Matzel	Matzel	Matzel_battery_FC	Mus musculus	NOR	No	Temporal	no	notForR	Gaussian	1	0.859	NA	1	0.946	NA	NA	NA	NA
Matzel	Matzel	Matzel_battery_LM	Mus musculus	NOR	No	Temporal	no	notForR	Gaussian	1	0.525	NA	1	0.671	NA	NA	NA	NA
Matzel	Matzel	Matzel_battery_OD	Mus musculus	NOR	No	Temporal	no	notForR	Gaussian	1	0.433	NA	1	0.410	NA	NA	NA	NA
Matzel	Matzel	Matzel_battery_PA	Mus musculus	NOR	No	Temporal	no	notForR	Gaussian	1	0.055	NA	1	0.224	NA	NA	NA	NA
Matzel	Matzel	Matzel_battery_WM	Mus musculus	NOR	No	Temporal	no	notForR	Gaussian	1	0.572	NA	1	0.574	NA	NA	NA	NA
Matzel	Matzel	Matzel_attention_internal	Mus musculus	NOR	No	Contextual	no	notForR	Gaussian	1	0.794	NA	1	0.907	NA	NA	NA	NA
Matzel	Matzel	Matzel_attention_external	Mus musculus	NOR	No	Contextual	no	notForR	Gaussian	1	0.622	NA	1	0.454	NA	NA	NA	NA
Matzel	Matzel	Matzel_attention	Mus musculus	NOR	No	Contextual	no	notForR	Gaussian	1	0.646	NA	1	0.697	NA	NA	NA	NA
Nawroth	Nawroth	Nawroth_PDS	Sus scrofa	NBT	R Rn Rni	Temporal	no	notForR	Poisson	1	0.027	0.998	1	0.005	1.000	1	0.021	1.000
Nawroth	Langbein	Nawroth_goat_EP_indirect_visual	Capra hircus	ACC	No	Temporal	no	notForR	Gaussian	1	0.950	NA	1	0.549	NA	1	0.762	NA
Nawroth	Nawroth	Nawroth_byExp	Sus scrofa	NBT	No	Contextual	no	notForR	Gaussian	1	0.782	NA	1	0.564	NA	1	0.749	NA
Nawroth	Langbein	Nawroth_goat_EP_visual_acoustic	Capra hircus	ACC	No	Contextual	no	notForR	Gaussian	1	0.603	NA	1	0.845	NA	1	0.897	NA
Nawroth	Langbein	Nawroth_goat_OP	Capra hircus	ACC	No	Contextual	no	notForR	Gaussian	1	0.766	NA	1	0.988	NA	1	0.501	NA
Nawroth	Langbein	Nawroth_goat_HC	Capra hircus	ACC	No	Contextual	no	notForR	Gaussian	1	0.546	NA	1	0.711	NA	1	0.962	NA
Rodr1guez	Gloudeman	Rodr1guez_ME_3	Frontinella communis	LAT	NA	Contextual	yes	forR	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Rodriguez	Gloudeman	Rodriguez_within_1	Frontinella communis	LAT	NA	Temporal	yes	forR	Gui	NA	NA	NA	NA	NA	NA	NA	NA	NA
Rodriguez	Gloudeman	Rodriguez_within_2	Frontinella communis	LAT	NA	Temporal	yes	forR	Gui	NA	NA	NA	NA	NA	NA	NA	NA	NA
Rodriguez	Gloudeman	Rodriguez_within_3	Frontinella communis	LAT	NA	Temporal	yes	forR	Gui	NA	NA	NA	NA	NA	NA	NA	NA	NA
Rodriguez	Gloudeman	Rodriguez_within_4	Frontinella communis	LAT	NA	Temporal	yes	forR	Gui	NA	NA	NA	NA	NA	NA	NA	NA	NA
Rodriguez	Gloudeman	Rodriguez_within_5	Frontinella communis	LAT	NA	Temporal	yes	forR	Gui	NA	NA	NA	NA	NA	NA	NA	NA	NA
Rodriguez	Gloudeman	Rodriguez_within_6	Frontinella communis	LAT	NA	Temporal	yes	forR	Gui	NA	NA	NA	NA	NA	NA	NA	NA	NA
Rodriguez	Gloudeman	Rodriguez_ME_1	Frontinella communis	LAT	NA	Contextual	yes	forR	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Rodriguez	Gloudeman	Rodriguez_ME_2	Frontinella communis	LAT	NA	Contextual	yes	forR	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Schuster	Schuster	Schuster_within_1	Micromys minutus	LAT	NA	Temporal	yes	forR	Gui	NA	NA	NA	NA	NA	NA	NA	NA	NA
Schuster	Schuster	Schuster_within_2	Micromys minutus	TTC	NA	Temporal	yes	forR	Poi	NA	NA	NA	NA	NA	NA	NA	NA	NA
Schuster	Schuster	Schuster_within_3	Micromys minutus	PRO	NA	Temporal	yes	forR	Gui	NA	NA	NA	NA	NA	NA	NA	NA	NA
Schuster	Schuster	Schuster_within_4	Micromys minutus	LAT	NA	Temporal	yes	forR	Gui	NA	NA	NA	NA	NA	NA	NA	NA	NA
Schuster	Schuster	Schuster_within_5	Micromys minutus	LAT	NA	Temporal	yes	forR	Gui	0	NA	NA	NA	NA	NA	NA	NA	NA
Schuster	Schuster	Schuster_within_6	Micromys minutus	LAT	NA	Temporal	yes	forR	Gui	NA	NA	NA	NA	NA	NA	NA	NA	NA
Shaw	Shaw	Shaw_Motor	Petroica longipes	TTC	NA	Contextual	yes	forR	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Shaw	Shaw	Shaw_IN	Petroica longipes	TTC	NA	Contextual	yes	forR	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Sorato	Lovlie	Sorato_all	Gallus gallus	TTC	No	Contextual	no	notForR	Poisson	1	0.165	0.990	1	0.148	0.995	1	0.165	0.989
Van Horik	Emery	Van Horik_caicque_RL	Pionites melanocephala	TTC	No	Temporal	no	notForR	Gaussian	1	0.449	NA	1	0.890	NA	NA	NA	NA
Van Horik	Emery	Van Horik_Macaw_TTC	Diopsittaca nobilis	TTC	No	Temporal	no	notForR	Gaussian	1	0.727	NA	1	0.926	NA	1	0.947	NA
Van Horik	Madden	Van Horik_pheasant_TotalCorrect	Phasianus colchicus	NBT	No	Contextual	no	notForR	Gaussian	1	0.535	NA	1	0.679	NA	1	0.816	NA
Wilkinson	Wilkinson	Wilkinson_tortoise_reversal	Chelonoidis carbonaria	TTC	No	Temporal	no	unpub	Gaussian	1	0.564	NA	1	0.449	NA	NA	NA	NA
Wilkinson	Wilkinson	Wilkinson_tortoise_Learning	Chelonoidis carbonaria	TTC	No	Contextual	no	unpub	Gaussian	1	0.753	NA	1	0.910	NA	NA	NA	NA
Wilkinson	Wilkinson	Wilkinson_tortoise_reversal1	Chelonoidis carbonaria	TTC	No	Contextual	no	unpub	Gaussian	1	0.908	NA	1	0.691	NA	NA	NA	NA

Table S3: Detailed information on exclusion and analysis for each study. Residuals that were significantly ($p<0.05$) non uniform (uni) or overdispersed (over) were excluded from the analysis. R = unadjusted repeatability; Rn = repeatability adjusted for test order; Ri = repeatability adjusted for test order and individual determinants.

Mean effect size	Low CI	High CI	Mean effect size Phylo	Low CI Phylo	High CI Phylo
0.183	0.088	0.282	0.164	0.001	0.343
0.15	0.095	0.213	0.153	0.061	0.269
0.279	0.129	0.428	0.275	0.114	0.435
0.267	0.032	0.477	0.267	0.032	0.477
0.222	0.129	0.312	0.222	0.129	0.312
0.195	0.085	0.305	0.195	0.085	0.305

Table S4: Mean intercept and confidence interval of meta-analytic model with (Phylo) or without controlling for phylogenetic effect.

	Temporal R	Temporal R adjusted for test order	Temporal R adjusted for test order and individual determinants	Contextual R	Contextual R adjusted for test order	Contextual R adjusted for test order and individual determinants
Temporal R	NA	0.05	0.201	0.5	0.313	0.34
Temporal R adjusted for test order	NA	NA	0.315	0.127	0.009	0.018
Temporal R adjusted for test order and individual determinants	NA	NA	NA	0.26	0.101	0.12
Contextual R	NA	NA	NA	NA	0.377	0.388
Contextual R adjusted for test order	NA	NA	NA	NA	NA	0.491
Contextual R adjusted for test order and individual determinants	NA	NA	NA	NA	NA	NA

Table S5: P-value matrix resulting from multiple t-test comparisons for each combination of R analysis before correction for multiple comparisons.

Identification	Total number of repeatability studies searched using search term: 'Repeatability and Cognition'; 'individual consistency in cognitive performance'; 'repeatability of cognitive performance', ''rptR' package and cognitive performance' in Web of Science. **N = 379**	Total number of researchers contacted (in the workshop, via personal contact): **N=61 (36, 25)**
Screening	Number of studies reporting repeatability of cognitive performance (R): **N= 6**	Received primary data (from people in the workshop, from personal contact) **N = 38 (15, 23)**

Eligibility		R already published (papers)	R computed from primary data (datasets)
	Number of studies for which estimates of temporal repeatability were provided:	17 (3)	89 *(22)*
	Number of studies for which estimates of contextual repeatability were provided:	18 (5)	*89* *(27)*
Included	Number of published studies included in the meta-analysis: **N=6** Number of unpublished studies included in the meta-analysis: **N=38** **Total datasets: 44** **Total analyses: 213 analyses**		

Figure S1. PRISMA diagram for repeatability of cognitive performance meta-analysis.

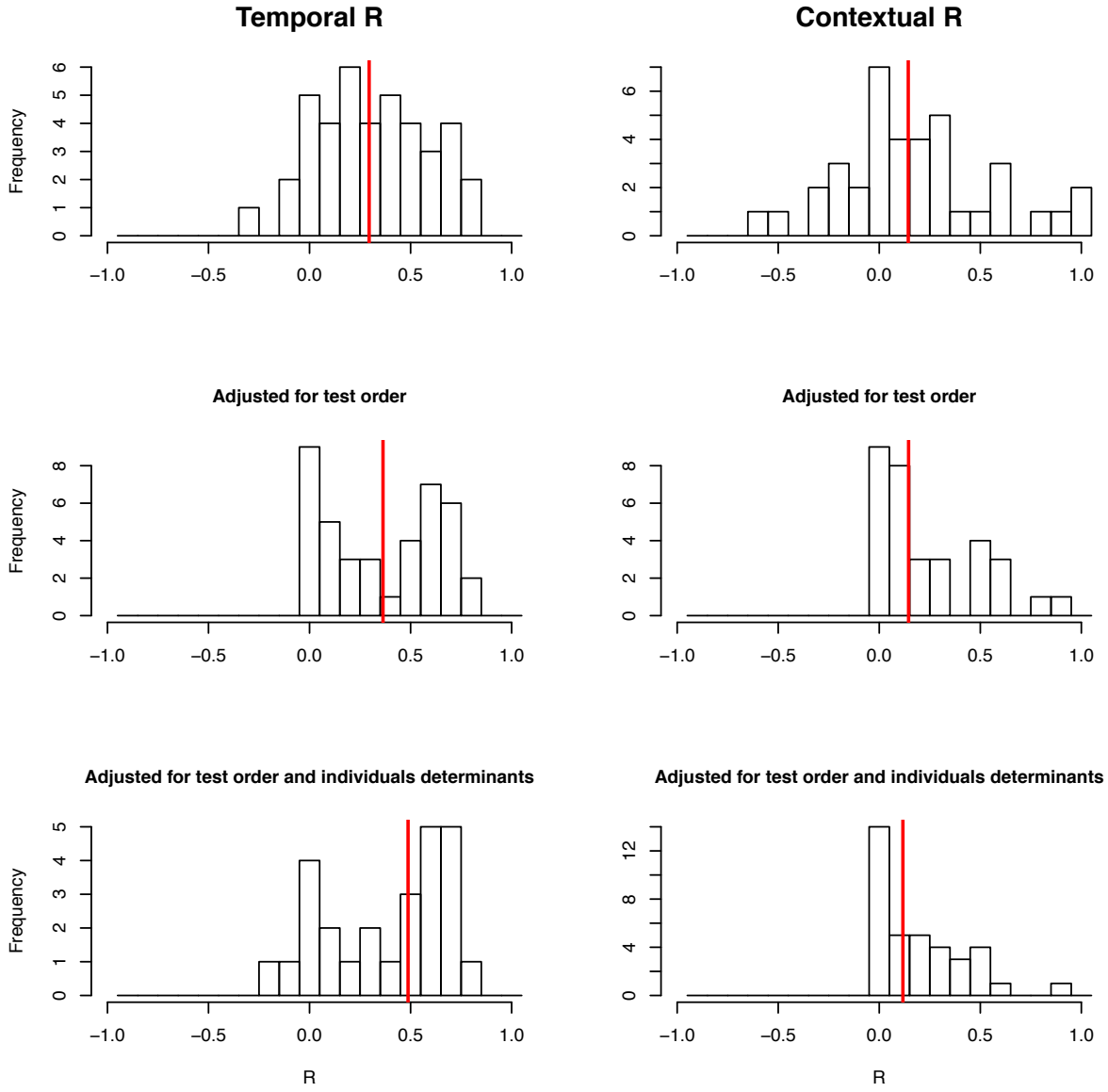


Figure S2: Distribution of repeatability values (R) for experiments with temporal repetition (Left) and contextual repetition (Right). Unadjusted R are represented on the top row. Adjusted R for test order are represented in the middle row. Adjusted R both for test order and individual determinants (sex and/or age) are represented on the bottom row. The vertical red line indicates median R.

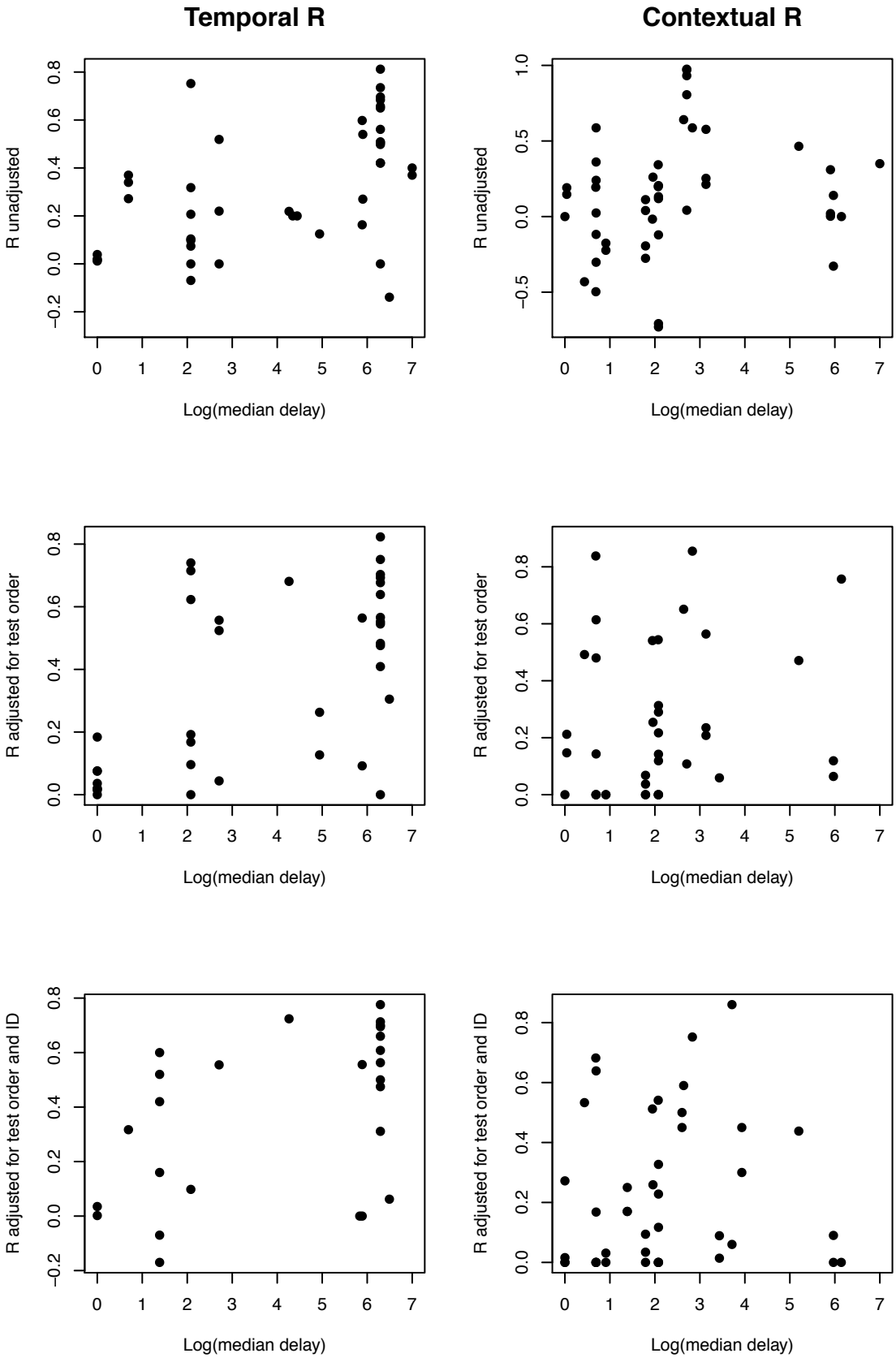


Figure S3: R according to median delay between two repeated test for temporal (left) and contextual (right) R. Unadjusted R are represented on the top row. Adjusted R for test order

are represented in the middle row. Adjusted R both for test order and individual determinants (sex and/or age) are represented on the bottom row.

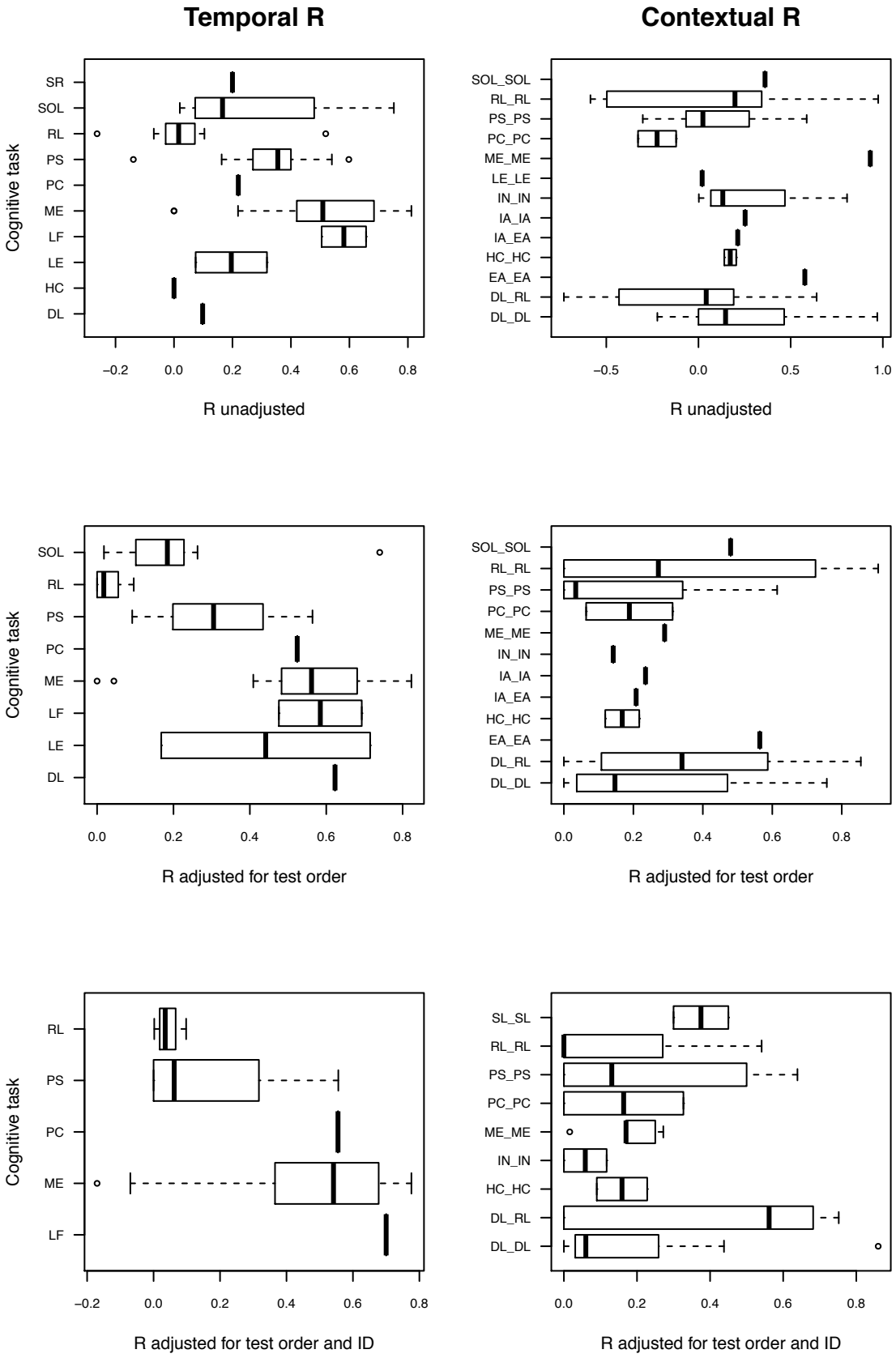


Figure S4: R according to cognitive task type for temporal (left) and contextual (right) R. Unadjusted R are represented on the top row. Adjusted R for test order are represented in the middle row. Adjusted R both for test order and individual determinants (sex and/or age) are represented on the bottom row

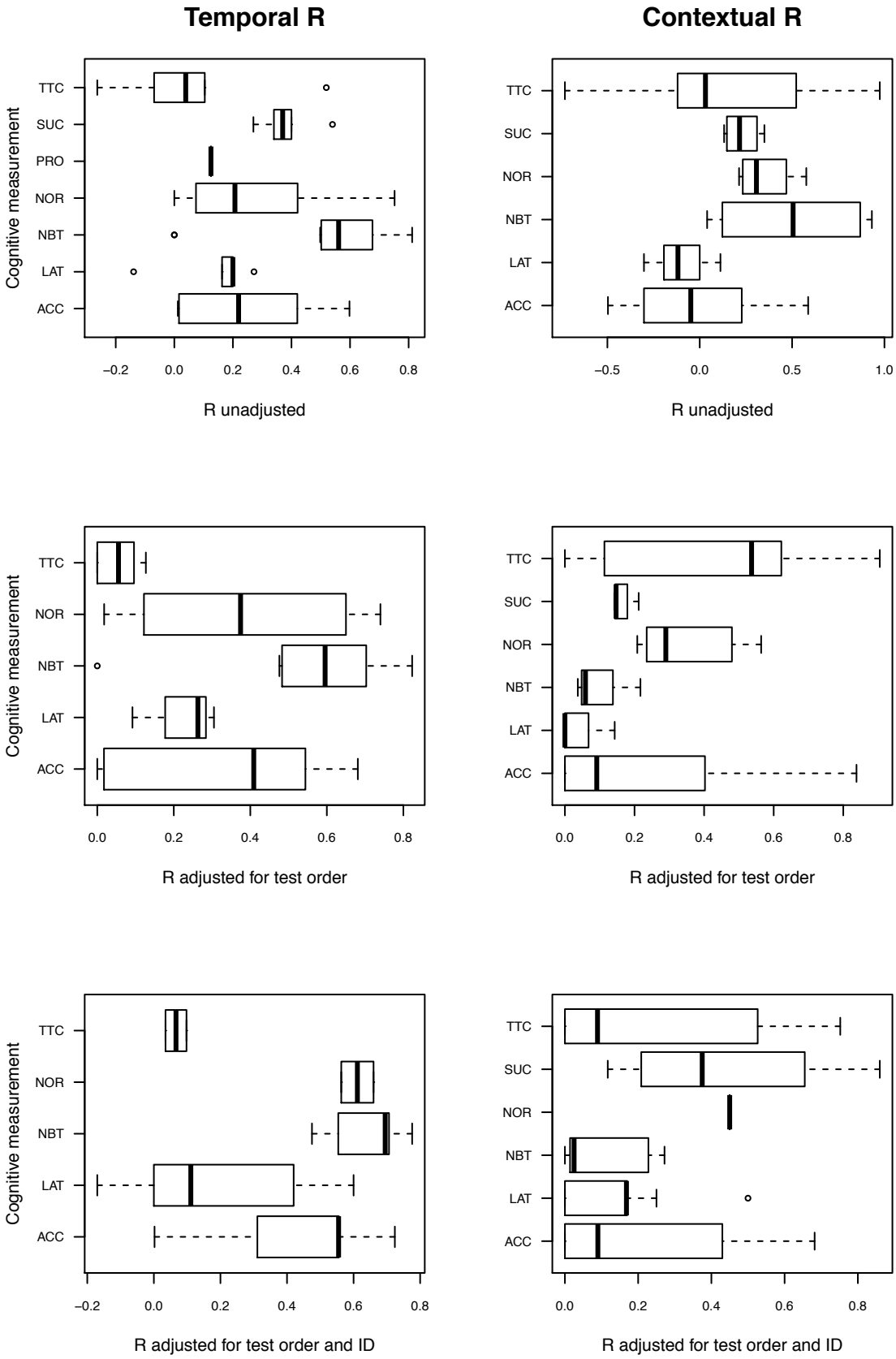


Figure S5: R according to cognitive performance measurement for temporal (left) and contextual (right) R. Unadjusted R are represented on the top row. Adjusted R for test order

are represented in the middle row. Adjusted R both for test order and individual determinants (sex and/or age) are represented on the bottom row.

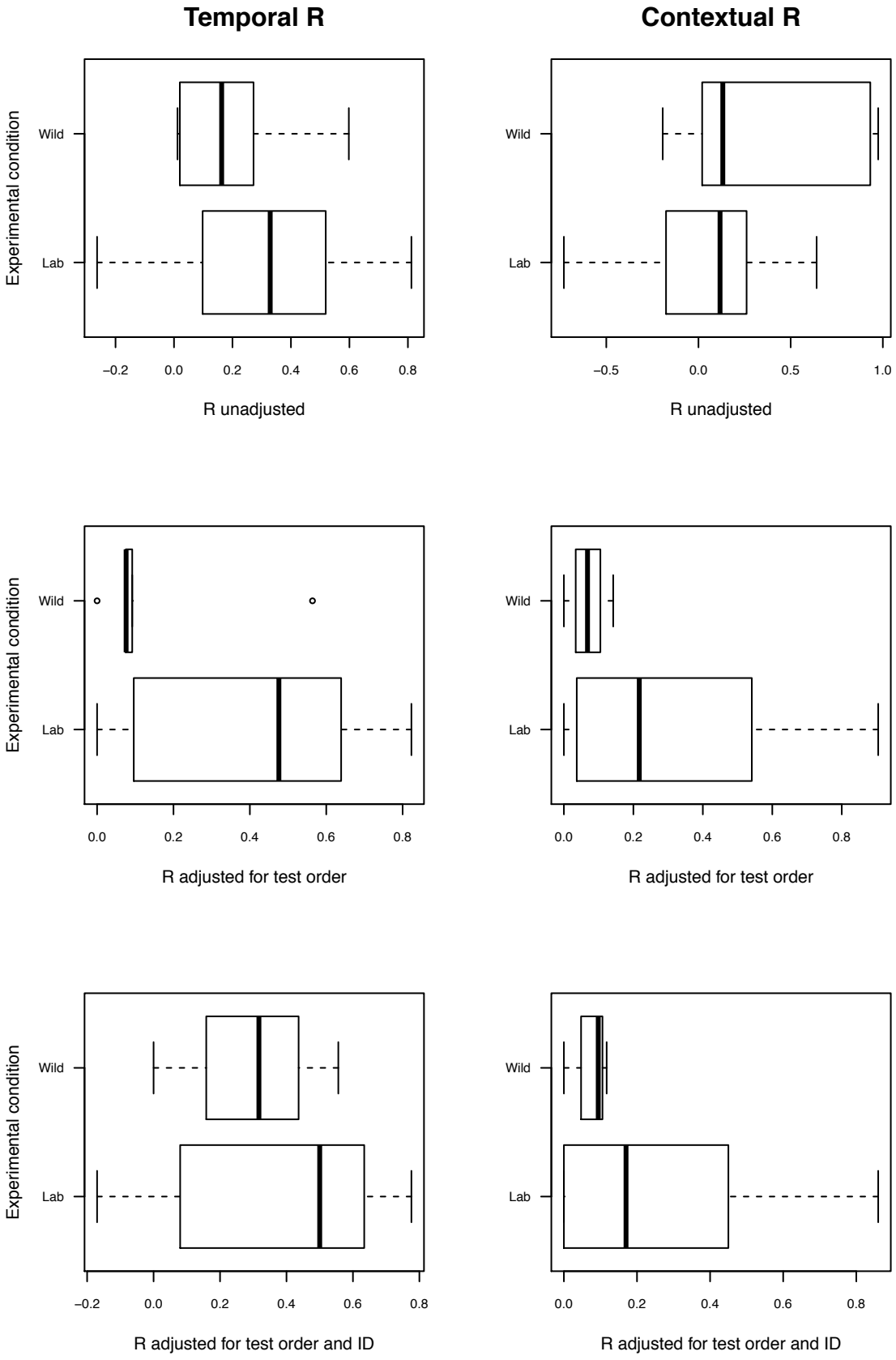


Figure S6: R according to experimental context for temporal (left) and contextual (right) R. Unadjusted R are represented on the top row. Adjusted R for test order are represented in the middle row. Adjusted R both for test order and individual determinants (sex and/or age) are represented on the bottom row.

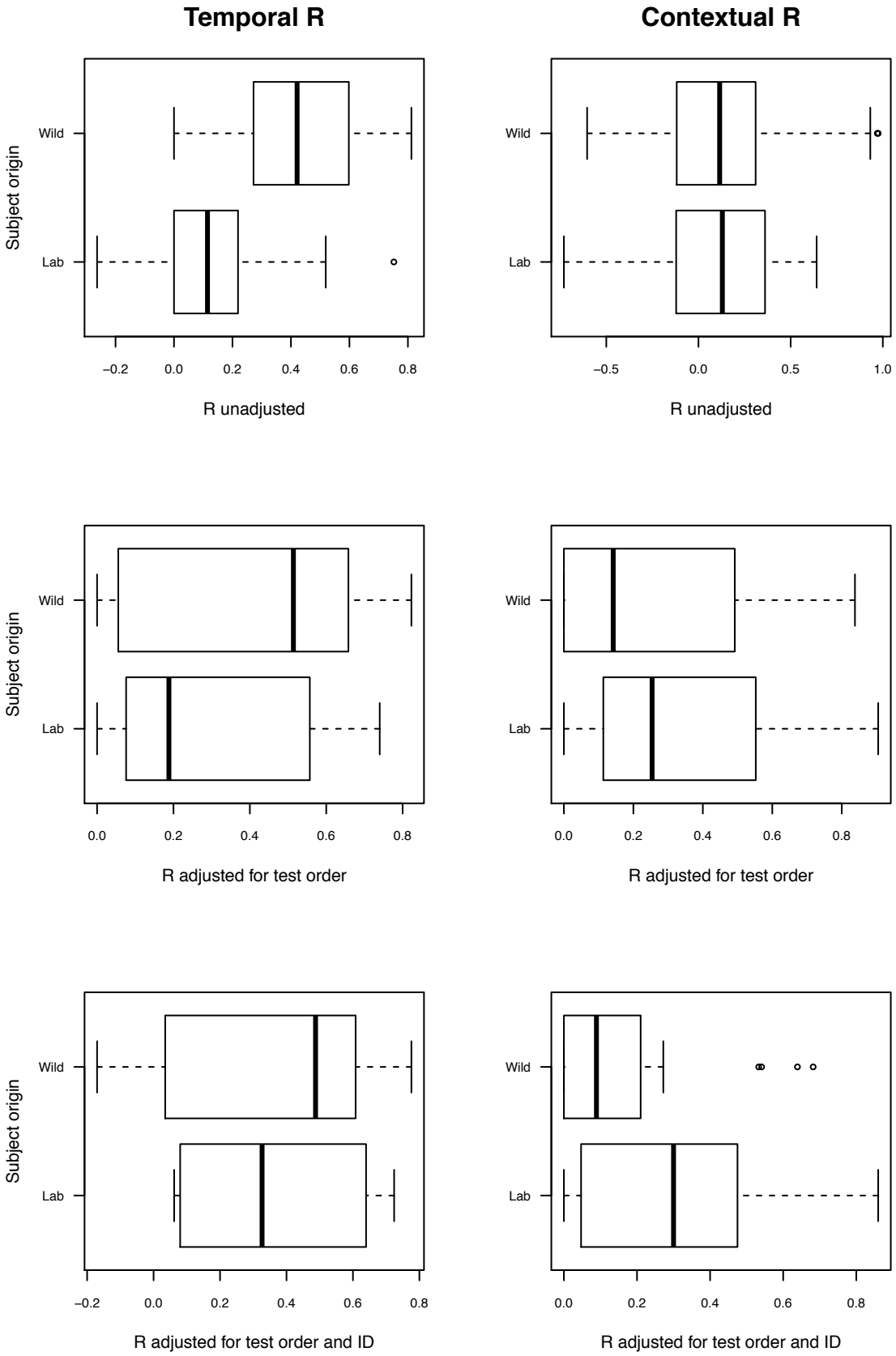


Figure S7: R according to subject origin for temporal (left) and contextual (right) R. Unadjusted R are represented on the top row. Adjusted R for test order are represented in the

middle row. Adjusted R both for test order and individual determinants (sex and/or age) are represented on the bottom row.

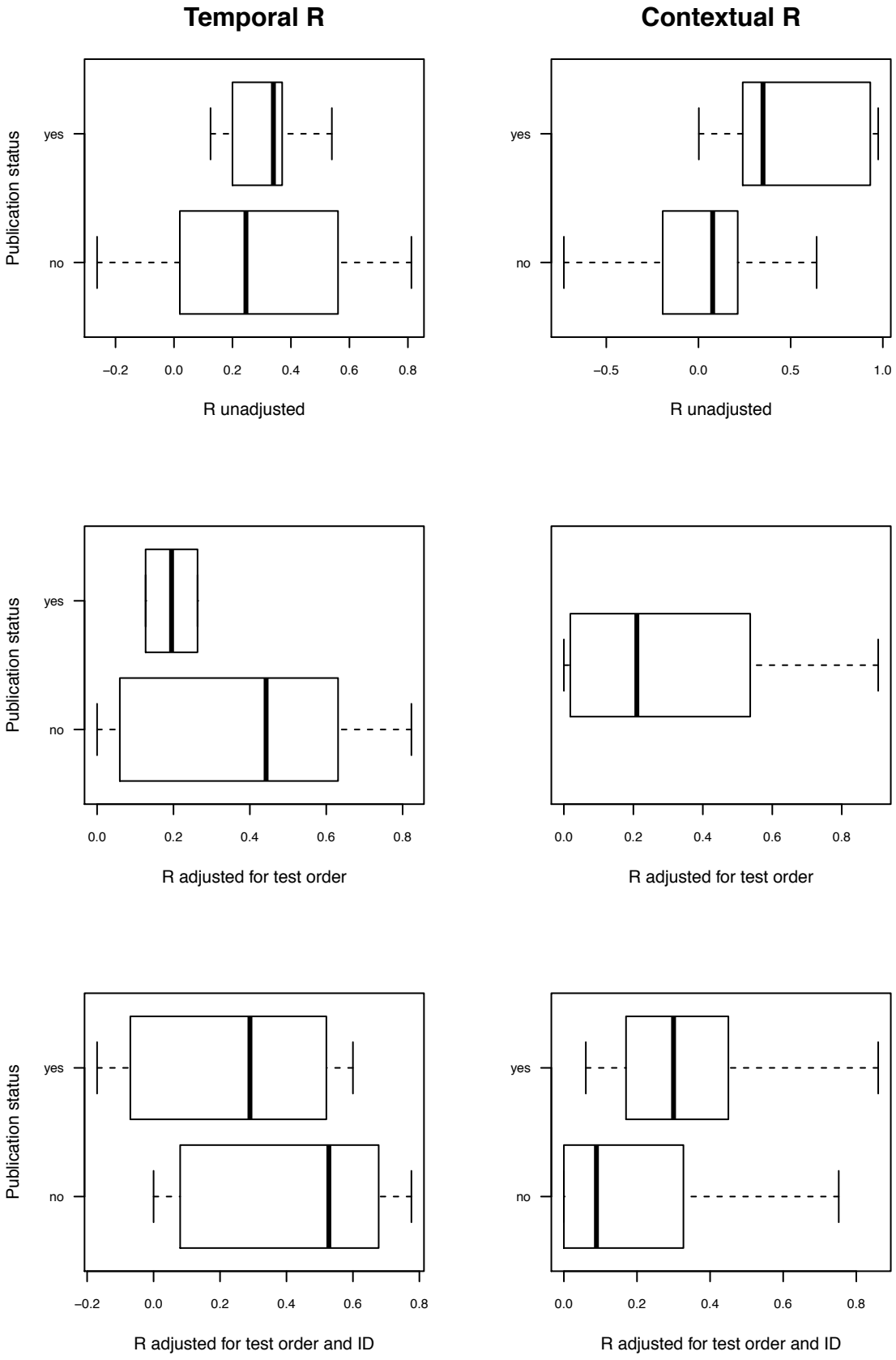


Figure S8: R according to publication status (yes= R published; no= R computed from primary data in the present paper) for temporal (left) and contextual (right) R. Unadjusted R are represented on the top row. Adjusted R for test order are represented in the middle row. Adjusted R both for test order and individual determinants (sex and/or age) are represented on the bottom row.

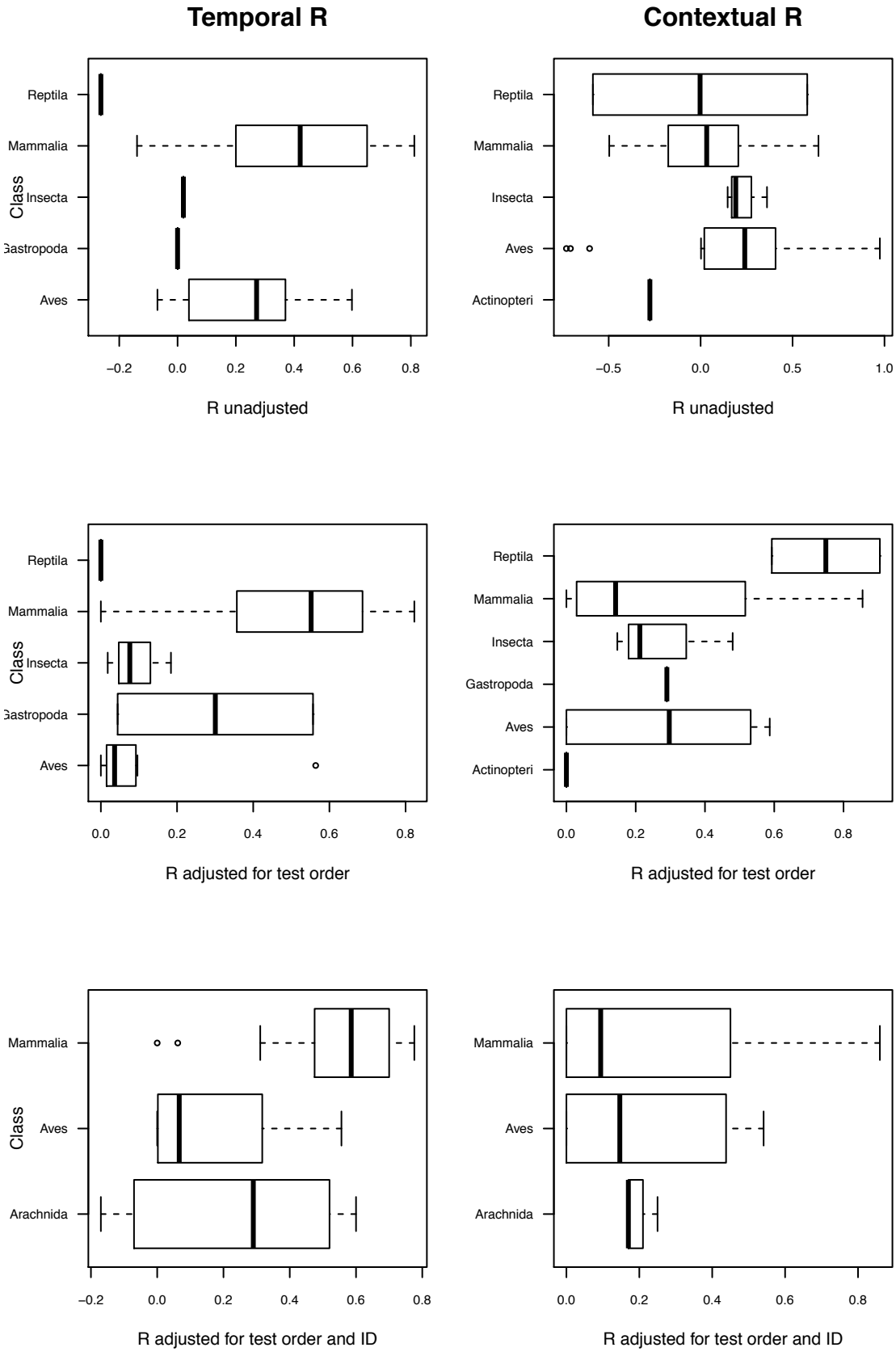


Figure S9: R according to taxonomic class for temporal (left) and contextual (right) R. Unadjusted R are represented on the top row. Adjusted R for test order are represented in the

middle row. Adjusted R both for test order and individual determinants (sex and/or age) are represented on the bottom row.

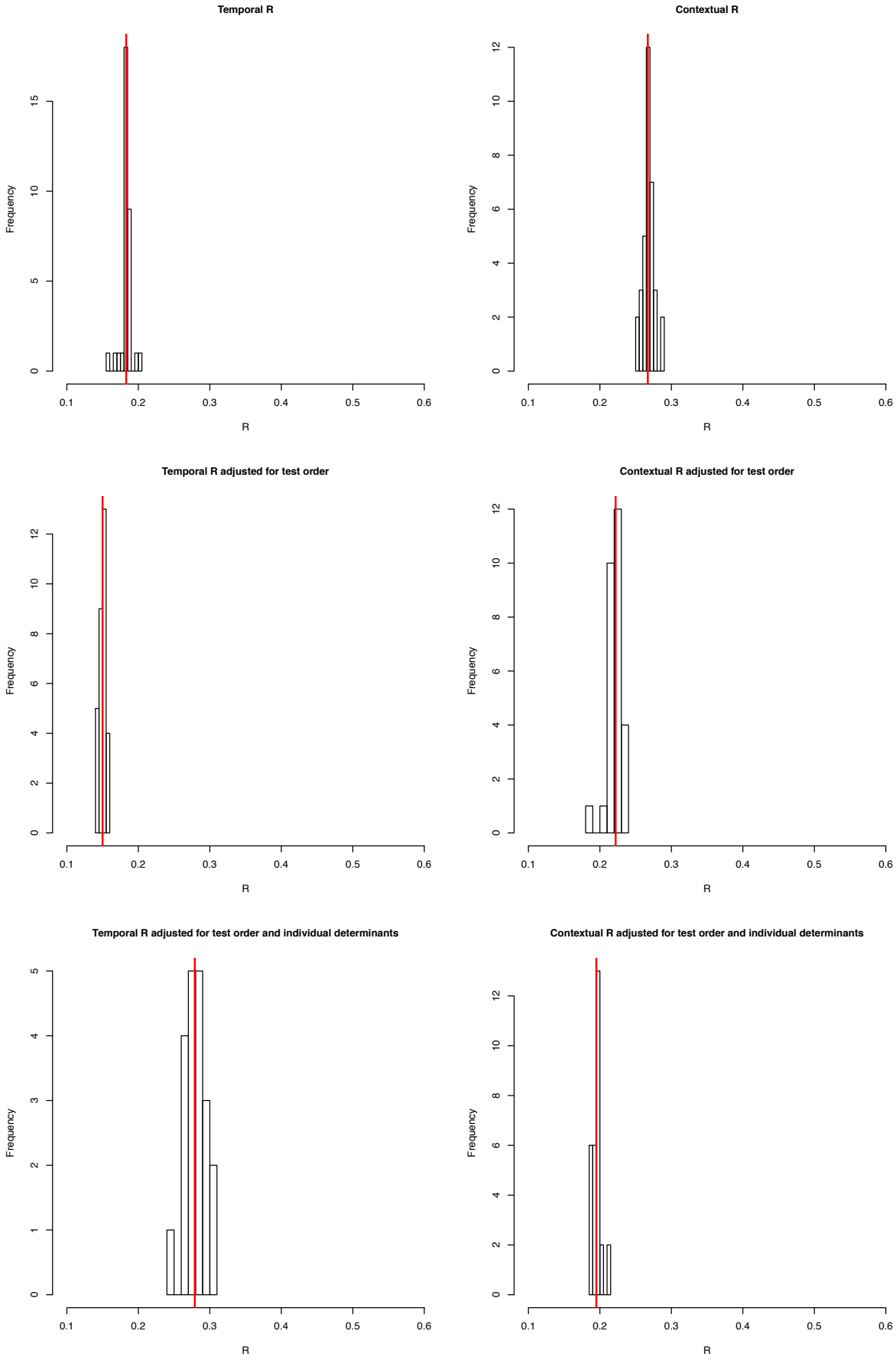


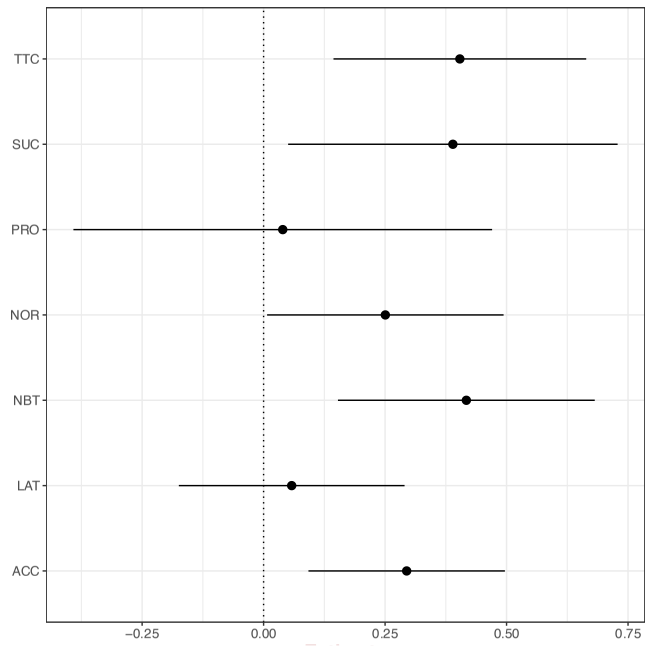
Figure S10: Frequency of R resulting from leave one out procedure for temporal (left) and

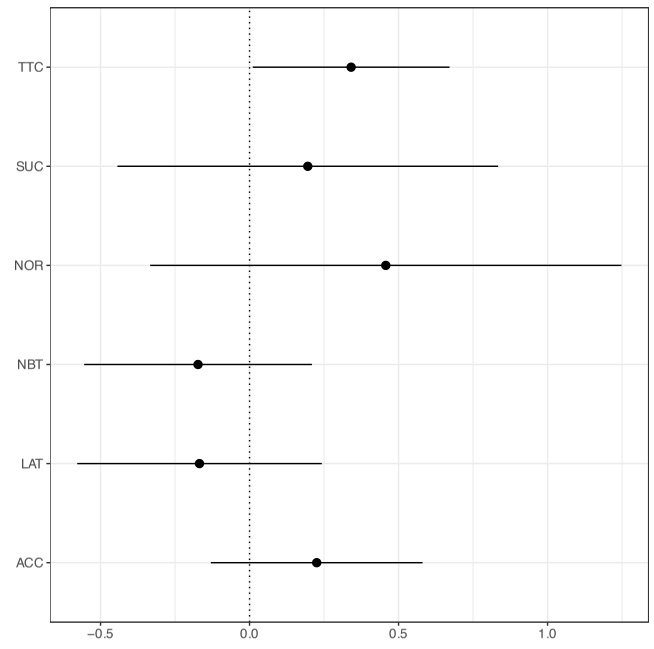
contextual (right) R. Vertical red line indicate mean estimate using all studies (Table 1). Unadjusted R are represented on the top row. Adjusted R for test order are represented in the middle row. Adjusted R both for test order and individual determinants (sex and/or age) are represented on the bottom row.

Temporal

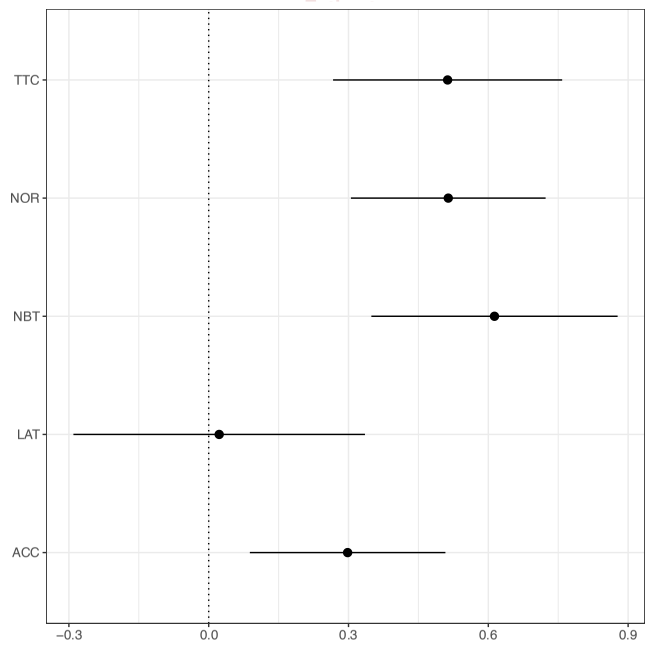
Contextual

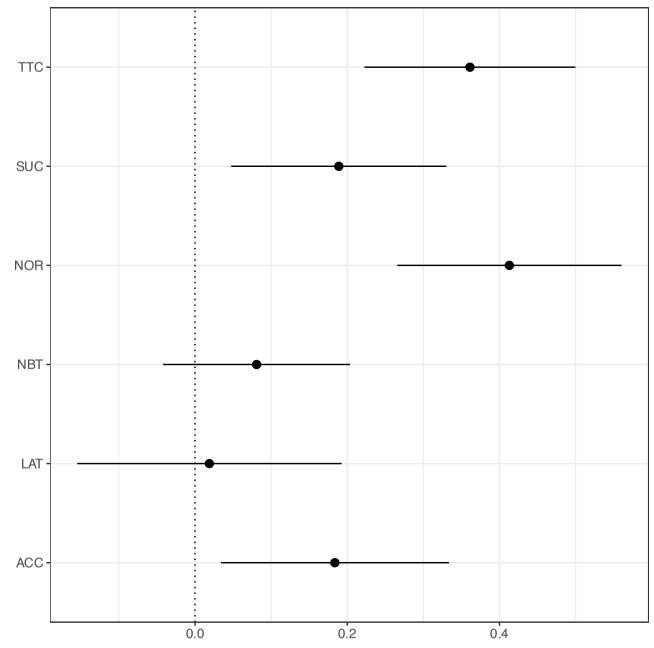
Unadjusted



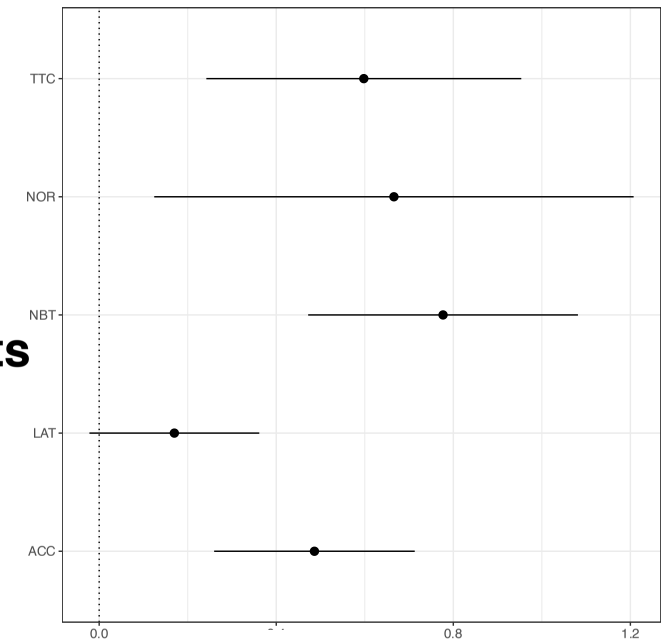


Adjusted for test order

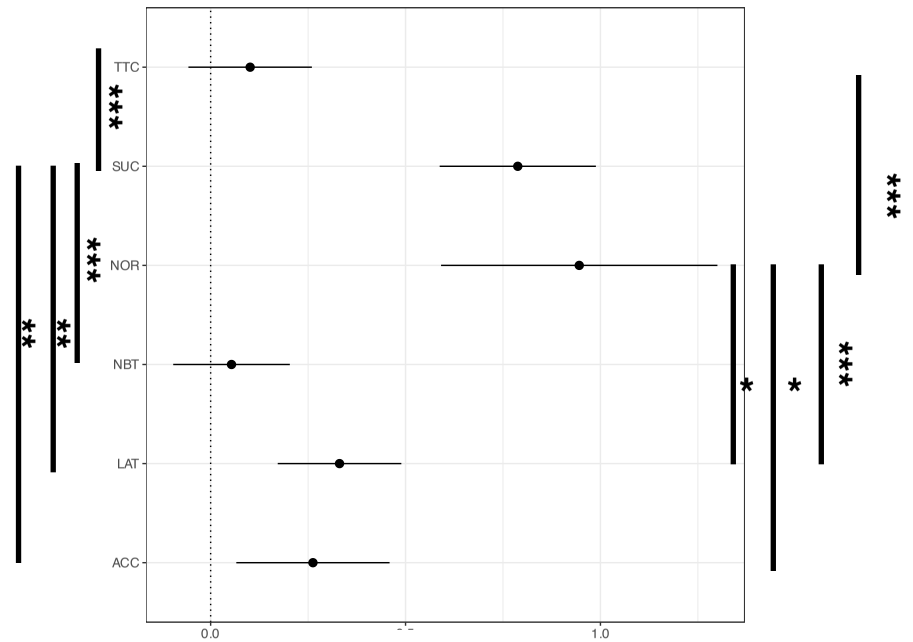




Adjusted for test order and individual determinants



β coefficient



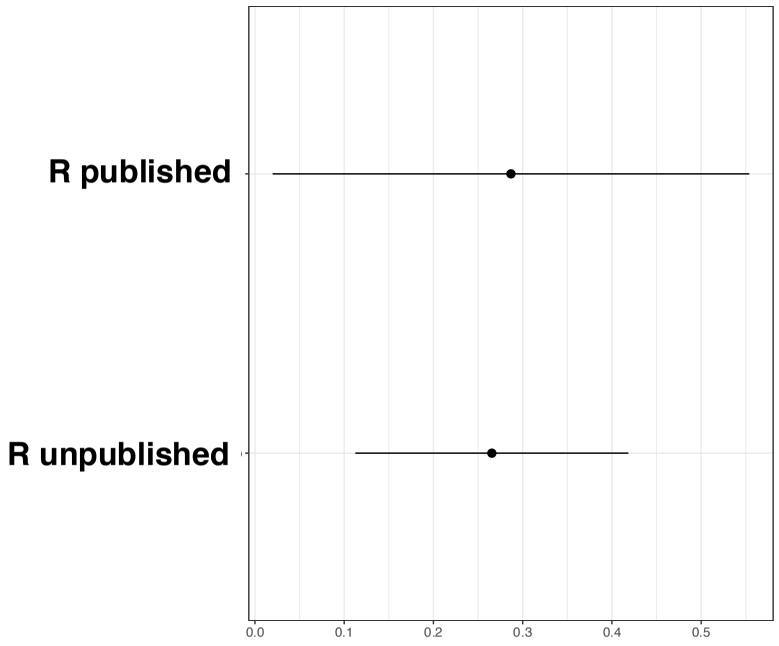
β coefficient

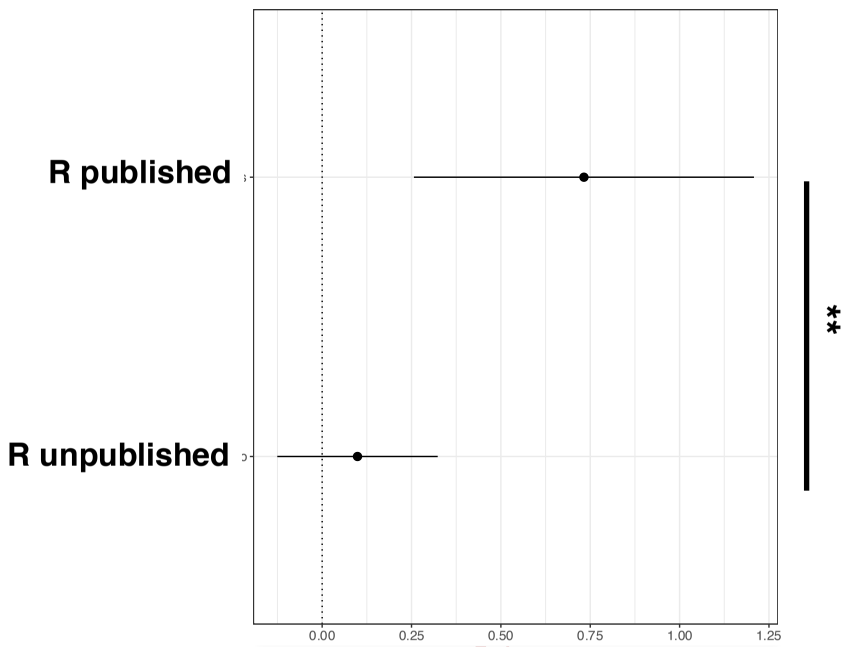
Figure S11: Beta coefficient (Fisher's Z normalized) of each cognitive measurement for temporal (left) and contextual (right) R. Unadjusted R are represented on the top row. Adjusted R for test order are represented in the middle row. Adjusted R both for test order and individual determinants (sex and/or age) are represented on the bottom row. Only significant differences are represented with *<0.05; **<0.01; ***<0.001.

Temporal

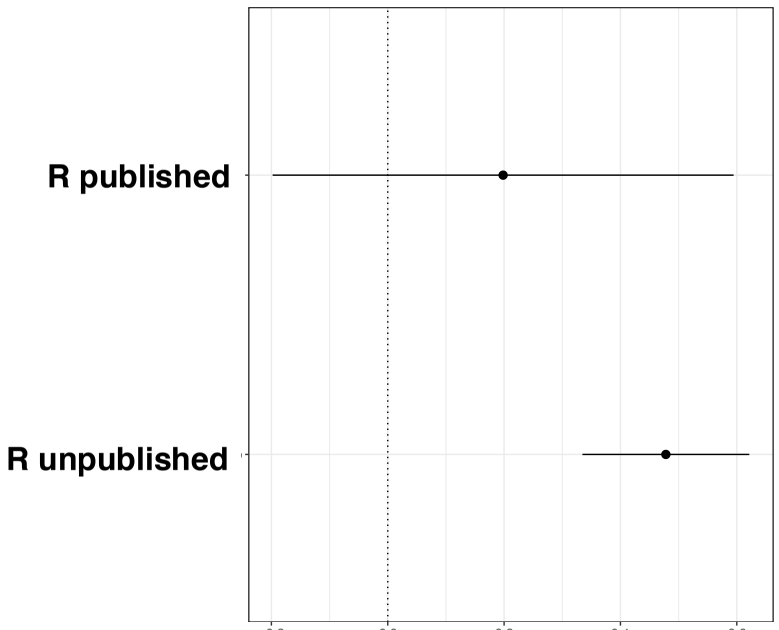
Contextual

Unadjusted



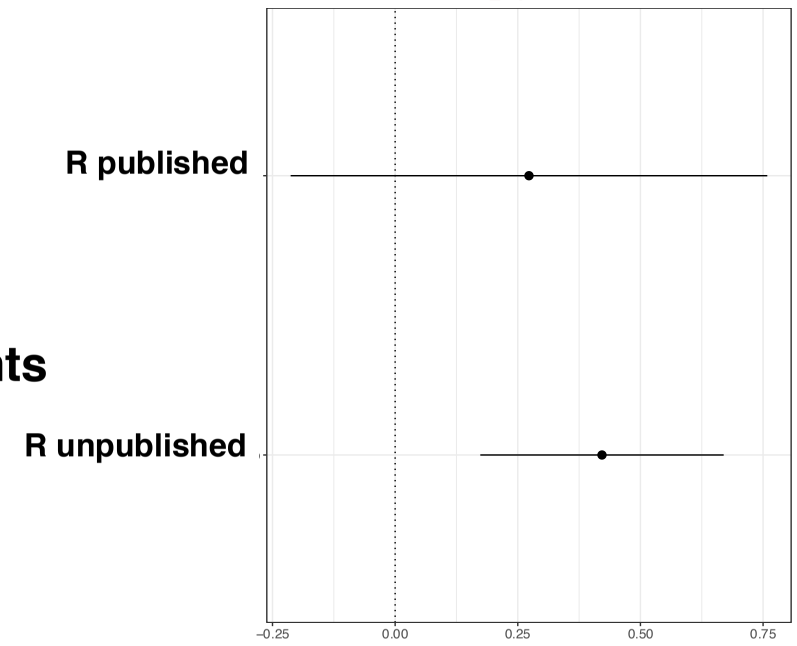


Adjusted for test order



NA

Adjusted for test order and individual determinants



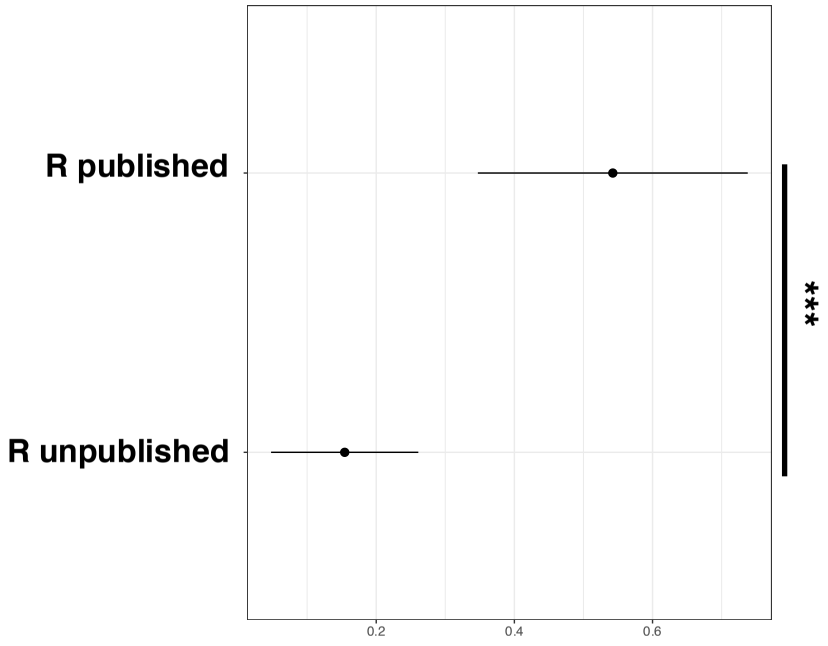


Figure S12: Beta coefficient (Fisher's Z normalized) for R published or computed from primary data (unpublished) for temporal (left) and contextual (right) R. Unadjusted R are represented on the top row. Adjusted R for test order are represented in the middle row. Adjusted R both for test order and individual determinants (sex and/or age) are represented on the bottom row. Only significant differences are represented with *<0.05; **<0.01; ***<0.001.