

How task format affects cognitive performance: a memory test with two species of New World monkeys

Schubiger, Michèle N.
Kissling, Alexandra
Burkart, Judith M.

This is the accepted manuscript © 2016, Elsevier
Licensed under the Creative Commons Attribution-NonCommercial-
NoDerivatives 4.0 International:

<http://creativecommons.org/licenses/by-nc-nd/4.0/>



The published article is available from doi:
10.1016/j.anbehav.2016.08.005

1 **How task format affects cognitive performance: a memory test with two species of New**
2 **World monkeys**

3

4

5 Michèle N. Schubiger^{1,2}, Alexandra Kissling¹ & Judith M. Burkart¹

6

7 ¹Department of Anthropology

8 University of Zurich

9 Winterthurerstrasse 190

10 CH-8057 Zurich

11 Switzerland

12

13 ²Division of Psychology

14 School of Social & Health Sciences

15 Abertay University

16 Kydd building

17 Dundee

18 Scotland

19

20 Corresponding author: Michèle N. Schubiger

21 michele.schubiger@uzh.ch/mnschubiger@gmail.com

22 Phone: +41 44 635 54 16

23 Fax: +41 44 635 68 04

24

25

26

27 **Highlights**

28

- 29 • Marmosets and squirrel monkeys were tested with two formats of a memory test.

30

- 31 • Performance was strongly affected by task format in both species.

32

- 33 • More options made random choices costly and increased the subjects' motivation.

34

35

- 36 • This finding has far-reaching consequences for comparisons within & across species.

37

38

39

40

41

42

43

44

45

46

47

48

49

50

51

52

53

54

55

56 **Abstract**

57 In cognitive tests, animals are often given a choice between two options and obtain a reward if
58 they choose correctly. We investigated whether task format affects subjects' performance in
59 two-choice cognition tests. In experiment 1, a 2-choice memory test, 15 marmosets (*Callithrix*
60 *jacchus*) had to remember the location of a food reward over time delays of increasing
61 duration. We predicted that their performance would decline with increasing delay, but this was
62 not found. One possible explanation was that the subjects were not sufficiently motivated to
63 choose correctly when presented with only two options because in each trial they had a 50%
64 chance of being rewarded. In experiment 2, we explored this possibility by testing naïve
65 marmosets (n = 8) and squirrel monkeys (*Saimiri sciureus*, n=7) with both the traditional two-
66 choice test *and* a new 9-choice version of the memory test that increased the cost of a wrong
67 choice. We found that task format affected the monkeys' performance. When choosing
68 between nine options, both species performed better and their performance declined as delays
69 became longer. Our results suggest that the 2-choice format compromises the assessment of
70 physical cognition, at least in memory tests with these New World monkeys, whereas providing
71 more options, which decreases the probability of obtaining a reward when making a random
72 guess, improves both performance and measurement validity of memory. Our findings suggest
73 that 2-choice tasks should be used with caution in comparisons within and across species
74 because they are prone to motivational biases.

75 **Keywords:** Marmosets, memory, physical cognition, squirrel monkeys, task format

76

77

78

79

80

81 **Introduction**

82

83 When the cognitive abilities of animals are assessed with cognitive tests, subjects are
84 often presented with two options to choose from and rewarded with a food item if they choose
85 the correct option. This two-choice task format has been used to test a variety of cognitive
86 abilities in a range of animal species such as memory (e.g. delayed response tasks in bees, *Apis*
87 *mellifera*; pigeons, *Columba livia*; several rat strains and many other species, including
88 primates; reviewed in Lind, Enquist, & Ghirlanda, 2015), understanding intentional deception
89 (chimpanzees, *Pan troglodytes*, Woodruff & Premack, 1979; and dogs, *Canis familiaris*, Petter,
90 Musolino, Roberts & Cole, 2009) and inferential reasoning (dogs, *Canis familiaris*, Erdőhegyi,
91 Topál, Virányi & Miklósi, 2007; birds, carrion crows, *Corvus corone corone*, Mikolatsch,
92 Kotrschal, & Schloegel, 2012; and primates, chimpanzees, *Pan troglodytes*, bonobos, *Pan*
93 *paniscus*; orangutans *Pongo pygmaeus*; and gorillas *Gorilla gorilla*, Call, 2006). One test that
94 has extensively used the two-choice format in particular with a wide range of animal species is
95 the object-choice task. The object-choice task tests for socio-cognitive abilities by assessing a
96 subject's ability to use an experimenter's gestural cues (e.g. gaze, point, touch) in order to
97 locate a reward that is hidden under one of usually two containers. The range of tested species
98 spans from primates (all four great apes and some Old and New World monkeys), domesticated
99 mammals (dogs, *Canis familiaris*; foxes, *Vulpes vulpes*; cats, *Felis catus*; horses, *Equus*
100 *caballus*; and goats, *Capra hircus*) and undomesticated terrestrial (wolves, *Canis lupus*; and
101 bats, *Pteropus spp.*) and marine mammals (dolphins, *Tursiops truncatus*; seals *Halichoerus*
102 *grypus* and *Arctocephalus pusillus*; and sea lions, *Otaria byronia*), to corvids (jackdaws,
103 *Corvus monedula*; and nutcrackers, *Nucifraga columbinana*) and parrots (African grey parrot,
104 *Psittacus erithacus*); see Mulcahy & Hedge (2012) for a review.

105 Although the two-choice task format is widely used in comparative psychology, there is
106 recent evidence that in some circumstances the task may not be a suitable method for assessing

107 cognitive abilities. Burkart & Heschl (2006), for instance, found that common marmosets
108 (*Callithrix jacchus*), a New World monkey species, chose at random when presented with only
109 two containers in an object-choice task, but they were able to use the experimenter's cues much
110 more reliably and made more correct choices when presented with nine instead of only two
111 containers to choose from. A likely explanation is that lowering the probability of obtaining a
112 reward by random choice helped the marmosets to overcome an inherent social bias that makes
113 non-human primates reluctant to follow communicative cues to food rewards.

114 In physical cognition tasks, such social biases should not influence a subject's
115 performance, because these tasks usually do not involve any social interaction between subject
116 and experimenter. Memory tests such as delayed response tasks (e.g. Kendrick, Rilling &
117 Denny, 1986; Lind et al., 2015; Rodriguez & Paule, 2009;) for instance, often require the
118 subjects to first observe and later remember in which of two locations a reward has been hidden
119 without obtaining any communicative cues. Consequently, if social biases alone were
120 responsible for the effect of task format on the marmosets' performance in the object choice
121 task, lowering the chance probability of success should not affect their performance in such
122 non-social cognition tasks. Nevertheless, the subjects may prefer to choose in a random manner
123 for other reasons, for instance to avoid the effort of memorizing. To date, it is not known if, or
124 to what extent, task format and chance probabilities also affect performance in physical
125 cognition tests. But if they do so in a similar way, as demonstrated for social tests, this has far-
126 reaching consequences for the validity of species comparisons that are often based on tasks that
127 differ in format.

128 In the present study, we tested New World monkeys with a physical cognition test that
129 assesses their memory ability and investigated if an alternative task format with nine choices
130 would also be more suitable than the traditional 2-choice task format. In experiment 1, we
131 tested common marmosets (*Callithrix jacchus*) with a traditional 2-choice memory test, i.e. the
132 memory subtest (hidden reward retrieval) of a cognitive test battery designed to assess general

133 intelligence in non-human primates (Banerjee et al., 2009). In this traditional delayed response
134 memory test, the subjects had to remember the location of a food reward over various time
135 delays. After watching how a food reward was hidden in one of two locations, the subject could
136 no longer see the reward and had to wait until the delay interval had expired before it could
137 choose one of the two locations. New World monkeys, particularly smaller species such as
138 marmosets (Miles, 1956; Miles, 1957a) and squirrel monkeys (French, 1959; Miles, 1957b,),
139 have been shown to perform worse on such delayed response tasks than Old World monkeys
140 (mainly rhesus macaques) and apes (e.g. Fischer & Kitchener, 1965; Harlow, 1932; Miles &
141 Meyer, 1956; reviewed in: Tomasello & Call, 1997). Even though the methodological details
142 are not always comparable, New World monkeys have also been shown to perform equally
143 well (capuchins, *Cebus apella*) or better (spider monkeys, *Ateles geoffroyi*) than Old World
144 monkeys (long-tailed macaques, *Macaca fascicularis*), and even comparable to great apes
145 (Amici, Aureli, and Call, 2010). Moreover, even smaller monkeys usually still perform well
146 above chance, at least with short delays (comparison of apes and monkeys, Fischer &
147 Kitchener, 1965). We therefore expected the marmosets to pass the traditional memory test in
148 experiment 1. Furthermore, in humans, success to remember a specific memory content
149 declines exponentially the more time has elapsed since its acquisition, a phenomenon known as
150 the forgetting curve (Ebbinghaus, 1885, 1913; hereafter Ebbinghaus effect). In experiment 1,
151 we therefore expected that the marmosets' performance would similarly decline with increasing
152 duration of the time delay if this test accurately measured memory performance. Since the
153 marmosets performed relatively poorly in experiment 1 and did not show an Ebbinghaus effect,
154 we conducted experiment 2. Experiment 2 was designed to assess the effect of reducing the
155 chance to obtain a reward when choosing at random. We tested a new sample of marmosets and
156 squirrel monkeys (*Saimiri sciureus*) and compared their performance in a traditional 2-choice
157 versus our newly developed 9-choice version of the memory test.

158

159 **Experiment 1: A traditional 2-choice memory test**

160

161 Methods

162 *Subjects*

163 Fifteen common marmosets (*Callithrix jacchus*), 8 males and 7 females participated in
164 this study. All subjects were housed in social groups consisting of two to six individuals at the
165 Primate Station of the Anthropological Institute and Museum of the University of Zurich. Their
166 indoor enclosures had both daylight and artificial light and were composed of one to three
167 components (depending on group size) measuring 1 m (width) x 2 m (depth) x 2 m (height),
168 each of which was equipped with several climbing structures such as natural branches, a
169 sleeping box, an infrared lamp, and a mulch floor. Whenever the weather conditions allowed it,
170 each group had free access to an outdoor enclosure. The marmosets were fed a vitamin and
171 calcium-enriched porridge in the morning, fresh fruit and vegetables at lunchtime, and gum and
172 mealworms in the late afternoon. In addition, they obtained a daily protein-snack in the
173 afternoon such as pieces of cooked egg. Water was available ad libitum from water dispensers.
174 All subjects were tested between their regular feedings and never food deprived during the
175 study. They could enter and leave the test enclosure through semi-transparent plastic tubes that
176 were connected to their home enclosures and were not handled at any time.

177

178 *Materials and Set-up*

179 Each subject was tested individually in the same compartment (41 cm x 53 cm x 33 cm)
180 of a larger test enclosure, with its group members present in an adjacent enclosure (100 cm x
181 122 cm x 78 cm) so that the subject could hear and smell, but not see them during testing. The
182 test compartment had a transparent Plexiglas window front containing two rectangular
183 openings (4 cm x 2.5 cm). The test apparatus consisted of two white opaque cylinder-shaped
184 plastic containers (3.0 cm in height and 5.3 cm in diameter) that were attached to a wooden

185 board (33 cm x 33 cm) placed 2 cm from its front, and was placed on the wooden testing table
186 (40 cm x 40 cm) that was level with the test compartment's floor. The test apparatus could be
187 slid in and out of the subject's reach. The two containers were filled with dark-brown bark
188 mulch that corresponded to the flooring substrate in the marmosets' home enclosures. A small
189 yellow piece of locust (*Schistocerca gregaria*) served as a reward in each trial. At the
190 beginning of each trial, the test apparatus was placed just out of the subject's reach and the two
191 containers were each covered with a rectangular mulch piece of approximately the same size as
192 the container.

193

194 *Procedure*

195 The experimenter stood behind the test apparatus, called the subject's name, said "look"
196 while showing it the reward and started a trial as soon as the subject was attentive. She
197 removed the cover of one of the two containers, placed the food reward in the container and
198 again covered it with the mulch piece so that the reward was no longer visible and both
199 containers, the baited and the empty one, remained covered. After the delay interval had
200 expired, she slid the board with the containers toward the test compartment's window. The
201 subject could then make a choice by reaching through one of two rectangular openings in the
202 window and removing the mulch cover with its hand(s). There were six delay conditions with
203 increasing time delays of 5, 10, 15, 20, 25, and 30 s. Each test session consisted of 10 trials of
204 one delay condition, if possible conducted on the same day, which resulted in a total of 60 trials
205 per subject. The reward's location was counter-balanced in a pseudo-randomized manner so
206 that a locust piece was hidden five times in the left and five times in the right container but
207 never in the same container in more than two consecutive trials. Prior to entering the actual test
208 sessions, each subject went through a pretest phase in which the experimenter followed the
209 same procedure but did not impose a time delay. After the subject reached criterion ($\geq 80\%$
210 correct choices within a single pretest session of 10 trials), it entered the test phase. At the

211 beginning of each testing day, the subject received one warm-up trial, again without a time
212 delay. Once a subject had finished the six test sessions, it was retested with one full session
213 without a delay. If the subjects had understood the task, we expected their performance in this
214 retest session to be higher than or at least as high as in the test sessions because the retest
215 involved no memory demand. We used two predefined criteria as to when to stop a test session:
216 1) the subject did not make a choice in three consecutive trials, and 2) the subject was no longer
217 attentive (not looking at the test apparatus but vigilant towards its surroundings instead) to the
218 task, and/or emotionally aroused (emitting vocalisations of discomfort and showing
219 piloerection of the tail; for definitions see Schubiger, Wüstholtz, Wunder, & Burkart, 2015), and
220 indicated it wanted to leave the test compartment (climbing to and rattling on the door on top of
221 the test compartment). If the subject met at least one of these criteria, it was allowed to go back
222 to its home enclosure and the session was continued the following day.

223

224 **Data scoring and analysis**

225 Of the 12 subjects who completed all test sessions, one male subject (Jugo) only
226 completed five trials of the retest and a second male subject (Vito) did not participate in the
227 retest. Three subjects, two males (Kapi and Kantor) and one female (Kitty), did not complete
228 the whole test phase, which resulted in a final total trial number of 756.

229 All trials were video recorded. The experimenter coded the subjects' choices live using
230 check sheets and checked all trials a second time using the video clips. Five trials (0.7%) had to
231 be excluded from the analysis owing to ambiguous behaviour of the subject or experimenter
232 error. A second rater coded 20 % of the trials from videos. The Kappa statistic was used to
233 determine the reliability between the two raters. Inter-rater reliability was excellent ($Kappa =$
234 $.96, P < 0.000, N=150$).

235 We ran a General Linear Mixed Model (GLMM) with delay condition as fixed and
236 subject as random factor to determine whether the delay condition significantly affected the

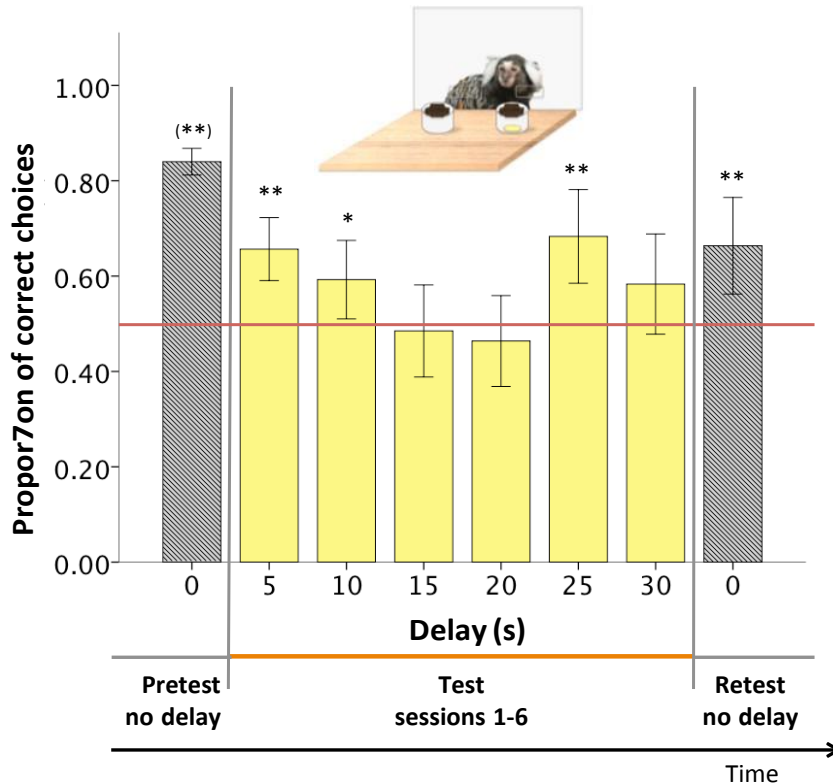
237 number of correct choices. Furthermore, we conducted one-sample t-tests to determine in
238 which of the six delay conditions the subjects performed above chance levels (more than 50%
239 correct choices) and whether their retest performance was still in the range of the criterion to
240 which they had been trained in the pretest phase.

241

242 **Results**

243 In the pretest phase, the marmosets reached criterion ($\geq 80\%$ correct choices within a
244 single session) within one to 11 sessions of 10 trials each (Mean = 2.93, SD = 2.55, $t_{14} = 3.70$,
245 $P = 0.002$). In the test phase, the marmosets chose the correct container across delay conditions
246 in 59% (SD = 8%) of all trials and thus significantly above chance ($t_{14} = 4.04$, $P = 0.001$). The
247 GLMM with delay condition as fixed factor and subject as random factor showed that the
248 duration of the delay had a significant effect on the subjects' performance ($F(5, 63.77) = 3.31$,
249 $P = 0.010$). We had also predicted that the subjects' performance in the test phase of
250 experiment 1 would decline with increasing length of the time delay, consistent with the
251 Ebbinghaus effect. However, after an initial decline of the number of correct choices that was
252 in line with this prediction, the subjects showed improved performance in the longest two delay
253 conditions (Fig. 1). A one-sample t-test demonstrated that the marmosets performed
254 significantly above chance after delays of 5 s (Mean = 66%, SD = 12%, $t_{14} = 5.12$, $P = 0.000$),
255 10 s (Mean = 59%, SD = 15%, $t_{14} = 2.42$, $P = 0.030$), and 25 s (Mean = 68%, SD = 16%, $t_{11} =$
256 4.01 , $P = 0.002$), but not after delays of 15 s (Mean = 48%, SD = 17%, $t_{13} = -0.34$, $P = 0.741$),
257 20 s (Mean = 46%, SD = 16%, $t_{12} = -0.81$, $p = 0.432$), and 30 s (Mean = 58%, SD = 17%, $t_{11} =$
258 1.70 , $P = 0.117$).

259 In the retest no-delay condition, the marmosets chose the correct container in 66 % of
260 all trials (SD = 16%), which is significantly above chance ($t_{10} = 3.46$, $P = 0.006$) and higher
261 than in five of the six test conditions, but differs significantly from the initial 80% criterion in
262 the pretest (Mean = 83%, SD = 5%); $t_{10} = 3.33$, $P = 0.008$).



264

265 **Figure 1** Performance in experiment 1. Subjects had to reach criterion ($\geq 80\%$ correct trials
 266 within a single session) in the pretest phase (no delay) before entering the test phase (delays = 5
 267 to 30 s) and were retested without a delay after completing the test phase. The red line indicates
 268 the chance level of 50%. Significance levels for above chance performance are indicated by * P
 269 $< .05$, ** $P < .01$. Error bars: 95% confidence intervals.

270

271 Discussion

272

273

274

275

276

277

278

279

280

281

We tested 15 common marmosets (*Callithrix jacchus*) with a traditional memory test (Banerjee et al, 2009). In this memory test, the subjects had to remember, over several time delay intervals ranging from five to 30 seconds, in which of two locations the experimenter had hidden a reward. All subjects passed the pretest phase, in which no delay was imposed, and, as a group, the marmosets also passed the test phase, by overall performing above chance. In contrast to our predictions, however, the marmosets' performance in the memory test did not decline with increasing delay duration, and they showed quite low levels of correct performance. It is unlikely that the marmosets were unable to remember the reward's location since they performed well after relatively long delays of up to half a minute. Moreover, saddle-back tamarins (*Saguinus fuscicollis*), another callitrichid species and close phylogenetic

282 relative, have been shown to remember the location of food items over much longer delay
283 intervals of up to 24 hours when tested in a naturalistic foraging task (Menzel, Juno and
284 Garrod, 1985). An alternative explanation for the marmosets' unexpected performance in the
285 hidden reward retrieval test (experiment 1) is that they may not have been sufficiently
286 motivated to choose correctly, particularly after short delays, because of the low cost of a
287 wrong choice. When choosing randomly between the two possible reward locations, they still
288 had a 50% chance of receiving a reward in each trial, and it was only after longer delays
289 between the experimenter's hiding action and the subject's choice that the cost of a wrong
290 choice increased owing to the longer waiting period.

291 We therefore designed a second experiment to explore if the task format, i.e. the
292 number of choice options, could explain the unexpected pattern of results in the traditional
293 memory subtest. Based on the findings of Burkart & Heschl (2006) in a modified object choice
294 task and our results from experiment 1, we developed a new memory test consisting of nine
295 choice options. This reduced the probability of making a correct choice by chance from 50% in
296 the 2-choice memory test to 11% and thus made a subject's wrong choice more costly. We
297 investigated if this 9-choice format, which had been shown to increase the performance of
298 marmosets in the above mentioned social cognition task, would also be more suitable than the
299 2-choice format in physical cognition tests. In order to do so, we compared the performance of
300 a naïve marmoset group in the traditional and our new memory test. In addition, we also tested
301 a group of squirrel monkeys (*Saimiri sciureus*), with the same two task formats and directly
302 compared the performance of the two species. This allowed us to evaluate whether task format
303 effects are specific to common marmosets or also present in other non-human primates. We
304 expected both species to perform better in the 9-choice memory test. Furthermore, we expected
305 the squirrel monkeys to outperform the marmosets as in previous delayed response studies
306 (Miles & Meyer, 1956; Miles, 1957b; Treichler, 1964; Tsujimoto & Savaguchi, 2002), owing
307 to their larger absolute and relative brain size (in proportion to their small body size, squirrel

308 monkeys have the largest brains of all primates; Rowe, 1996), which correlates with general
309 performance in physical cognition tasks (Deaner, van Schaik, & Johnson, 2006; Reader, Hager,
310 & Laland, 2011).

311

312 **Experiment 2: Introducing a new memory test format**

313 **Methods**

314 *Subjects*

315 Eight naïve common marmosets (*Callithrix jacchus*), four females and four males, and
316 seven male common squirrel monkeys (*Saimiri sciureus*) who had previous experience with a
317 similar test, participated in this study.

318 The housing conditions and feeding schedule of the marmosets corresponded to the
319 ones in experiment 1. The squirrel monkeys were housed in two bachelor groups of 3 and 5
320 individuals, respectively. Their indoor enclosures measured 16.55 m³ (smaller group) and 24.77
321 m³ (larger group) and were equipped with climbing structures, an infrared lamp, and a mulch
322 bark floor. The squirrel monkeys were fed a mixture of pellets and cottage cheese in the
323 morning, a variety of vegetables and a small amount of fruit at lunchtime, and a protein snack
324 such as cockroaches in the late afternoon. Since their indoor enclosures only had artificial UV-
325 light, each group had constant access to a fully roofed outdoor enclosure, and in addition, the
326 two groups took turns in accessing a larger outdoor area of 86.4 m³ whenever the weather
327 conditions allowed it. The squirrel monkeys could freely travel to and from the test enclosure
328 through a system of semi-transparent plastic tubes that connected it to their home enclosures.

329

330 *Set-up*

331 All subjects of both species were tested individually in a separate test compartment of a
332 larger test enclosure. The measurements of the marmosets' test compartment closely resembled
333 the ones in experiment 1, whereas the squirrel monkeys' test compartment measured 110 cm x

334 98 cm x 77 cm. We again used a test apparatus that could be slid forwards and backwards on a
335 testing table. The apparatuses for the marmosets (M) and the squirrel monkeys (S) were
336 identical and differed only in measurements that were adjusted to the marmosets' smaller body
337 size. It consisted of a wooden frame (M: 40 cm x 37.5 cm/S: 80 cm x 75 cm) containing three
338 wooden platforms (vertical distance between platforms M: 12.5 cm/S: 35 cm) that was
339 mounted on a wooden sliding board (M: 45 cm x 30 cm/S: 95 cm x 50 cm). Empty cylindrical
340 black plastic cups (diameter: 3.1 cm, height: M: 1.1 cm/S: 2.3 cm) with lids were used to hide
341 the reward. For the 9-choice test, three cups were placed equidistant (M: 14 cm/S: 29 cm)
342 between each outer and the middle cup) on each platform (outer cups at M: 4.5 cm/S: 11 cm
343 from the lateral frame). For the 2-choice test, 2 cups were placed on the middle platform (with
344 an in-between distance of M: 11 cm/S: 25 cm and at M: 10 cm/S: 25 cm distance from the
345 lateral frames). In both tests, the cups were held in place by Velcro tape strips. The front of the
346 test enclosure consisted of a lattice that allowed the subjects to reach out and choose one of the
347 cups.



348 **Figure 2** Test apparatus in experiment 2. Shown are both tests: (A) 2-choice, and (B) 9-choice
349 task format. (*Not drawn to scale; note: the lateral parts of the frame were solid boards*).

350

351

352 *Procedure*

353 The experimenter's procedure in the pretest and test phase corresponded to the one used
354 in experiment 1 with the exception of two additions in the test phase: 1) The experimenter said
355

356 “come” while pushing the apparatus toward the subject once the delay had expired in order to
357 encourage the subject to make its choice, and 2) the subject received one to three warm-up
358 trials (no delay) prior to each test session, and the test session only started once it had chosen
359 correctly in a warm-up trial. There were four increasing delay conditions ranging from 5 to 20
360 seconds and each test session consisted of 12 trials of one delay condition. When choosing
361 correctly, the subjects received their favourite rewards, mealworms or cashew nuts (squirrel
362 monkeys) and crickets or cooked apples (marmosets). The same stop criteria as in experiment 1
363 were used to decide when to terminate a session and continue testing on the next day.

364 We used a within-subject design in which every subject of each species was tested with
365 both task formats - the one with two choice options and the one with nine choice options, in
366 counterbalanced order. This resulted in two groups within each species: one group first
367 completed the 2-choice format followed by the 9-choice format while the second group was
368 tested in the opposite order. One male marmoset (Lexus) completed the whole 2-choice
369 memory test but only the 5-s delay condition in the 9-choice memory test. The final sample size
370 therefore consisted of eight marmosets (four females and four males) in the 2-choice and seven
371 marmosets (four females and three males) who completed all conditions in the 9-choice test, as
372 well as seven male squirrel monkeys, who completed both tests.

373

374 **Data scoring and Analysis**

375 All trials were video-recorded and the experimenter coded the subjects' choices live
376 using check sheets. A second rater coded 21 % of the trials from videos. The Kappa statistic
377 was used to determine the reliability between the two raters. For the marmosets one trial had to
378 be excluded owing to experimenter error (no delay imposed). Inter-rater reliability was
379 excellent (100%) for both squirrel monkeys ($Kappa = 1.00, P < 0.001, N=144$) and marmosets
380 ($Kappa = 1.00, P < 0.001, N=156$).

381 In order to test which factors best explained the subjects' performance, we ran
382 Generalized Linear Mixed Models (GLMMs) using the Restricted Maximum Likelihood
383 method (REML), with the fixed factors task format, species, delay condition, test order and
384 interactions (species*task format, species*delay, species*order, task format*delay, task
385 format*order, delay*order), and included subject as random factor. The best model was
386 determined using the AICc, the Akaike information criterion corrected for small sample sizes
387 (Hurvich & Tsai, 1989).

388 Since the probability to be successful by chance differed between the 2-choice and 9-
389 choice format, we could not use the subjects' raw scores to directly compare their performance
390 in the two conditions in the same statistical model but first had to compute a performance
391 measure that was independent of the task format. For this purpose, we computed a performance
392 score for each individual and condition according to the equation below, which corresponds to
393 the square root of the Chi-square value and in which a higher value represents better
394 performance. Observed values correspond to the individual number of correct choices per delay
395 (raw scores of 1 to 12) and expected values were calculated as the number of correct choices
396 expected by chance (6 out of 12 in the 2-choice and 1.33 out of 12 in the 9-choice memory
397 test).

$$\text{Performance score} = \sqrt{\frac{\text{observed} - \text{expected}}{\text{expected}}}$$

398 Finally, we conducted one-sample t-tests for each test format to determine in which conditions
399 the subjects performed above chance.

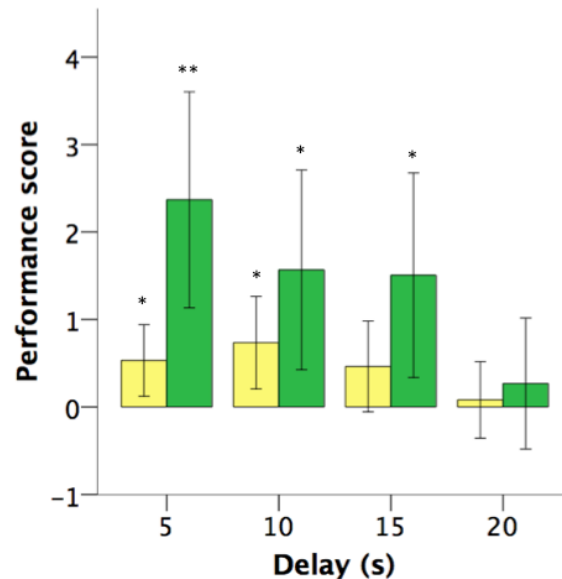
400

401 **Results**

402 In the pretest phase, the subjects reached criterion ($\geq 80\%$ correct within a single
403 session) after one to two sessions (Mean = 1.07; SD = 0.26) in the 2-choice memory test and
404 after one to seven sessions (Mean = 2.27; SD = 1.71) in the 9-choice task. The subjects took
405 significantly longer to reach criterion in the 9-choice than the 2-choice task: $t_{(14)} = -2.61$, $P =$

406 0.021. There was neither a species-difference in the number of pretest sessions in the 2-choice
407 (squirrel monkeys: Mean = 1.00, SD = 0.00; marmosets: Mean = 1.13, SD = 0.35; $t_{13} = -0.93$,
408 $P = 0.369$) nor in the 9-choice memory test (squirrel monkeys: Mean = 1.57; SD = 1.57;
409 marmosets: Mean = 2.88; SD = 2.10; $t_{13} = -1.54$, $P = 0.015$).

410 The best model included only the fixed effects test format and delay condition and no
411 interactions. Task format had a highly significant effect on the subjects' performance ($F_{1, 98.98} =$
412 18.13 , $P < 0.0001$) and so did delay condition ($F_{1, 98.29} = 5.65$, $P = 0.0013$). There was no
413 significant effect of species in any of the models (see table 2). Two separate GLMMs based on
414 raw scores of performance, one for each task format, with species, delay and order as fixed
415 factors and subject as random factor demonstrated that delay condition had a significant effect
416 on the subjects' per cent of correct choices for the 9-choice format ($F_{1, 39.88} = 5.46$, $P = 0.003$)
417 while there was only a trend for the 2-choice format ($F_{1, 42} = 2.49$, $P = 0.073$), see also Figure
418 3.



419

420 **Figure 3** The effect of task format on performance in experiment 2. The subjects' mean
421 performance (χ -transformed test scores to account for the two different chance levels) in the 2-
422 choice (yellow/light bars) and the 9-choice memory test (green/dark bars). *Error bars: 95%*
423 *confidence interval*. Asterisks * and ** indicate performance significantly above chance ($P <$
424 $.05$ and $P < .01$, respectively) in one-sample-t-tests on the raw values (percent correct choices).
425

426 In the 2-choice memory test, mean performance across all test sessions was
427 significantly above chance, i.e. > 50% correct choices (Mean = 59%, SD = 18%, $t_{59} = 4.02$, $P =$
428 0.000). Split-up per delay condition, the subjects as a group performed significantly above
429 chance in the 5-s (Mean = 61%, SD = 15%, $t_{14} = 2.87$, $P = 0.015$) and 10-s (Mean = 65%, SD =
430 19%, $t_{14} = 2.97$, $P = 0.010$) delay conditions, but not in the 15-s (Mean = 60%, SD = 19%, $t_{14} =$
431 1.92, $P = 0.076$) and 20-s conditions (Mean = 52%, SD = 16%, $t_{14} = 0.40$, $P = 0.695$). In the 9-
432 choice memory test, they also performed significantly above chance, i.e. > 11% correct
433 choices, across all test sessions (Mean = 25%, SD = 19%, $t_{56} = 5.43$, $P = 0.000$). Moreover,
434 they performed well above chance after delays of 5 s (Mean = 34%, SD = 21%, $t_{14} = 4.13$, P
435 = 0.001), 10 s (Mean = 26%, SD = 19%, $t_{13} = 3.00$, $P = 0.010$), 15 s (Mean = 26%, SD = 20%,
436 $t_{13} = 2.79$, $P = 0.015$), but not 20 s (Mean = 14%, SD = 13%, $t_{13} = 0.786$, $P = 0.446$).

437

438 **Discussion**

439 In experiment 2, we tested common marmosets and common squirrel monkeys, two
440 evolutionarily closely related species, with both the 2-choice and 9-choice task format of a
441 memory test. As predicted, we found that task format affected the performance of both species.
442 When the subjects were allowed to choose between nine rather than only two options, they
443 performed better, and, in line with our prediction, their performance decreased with longer
444 delays. However, the larger-brained squirrel monkeys did not outperform the marmosets,
445 although the small sample size makes it difficult to identify whether this finding is a true
446 absence of a species difference or instead reflects a lack of statistical power. In contrast to the
447 present findings, squirrel monkeys had outcompeted marmosets in delayed response studies.
448 However, some of these studies did not contain a true memory component (Miles & Meyer,
449 1956, Miles 1957b) or they contained a stronger working memory component (Tsujimoto &
450 Savaguchi, 2002). In the latter study, the reward's location was not randomized and subjects
451 had to keep in mind their previous choices and base their next choices strategically on these.

452 Both New World monkey species in our study remembered the reward's location for
453 longer time periods in the 9-choice format than in the 2-choice format. However, their
454 performance was still moderate, for instance in relation to closely related saddle-back tamarins
455 (Menzel et al., 1985) who remembered as many as 30 different locations over much longer
456 delays. Apart from species differences, a likely explanation is that the tamarin study was more
457 naturalistic than ours in that the subjects remained in the group setting during experiments, and
458 the locations to choose from were distributed over a much larger area. In fact, Fischer &
459 Kitchener (1965) had argued that delayed-response tasks with a strong spatial component are
460 easier to solve for non-human primates than non-spatial ones. The more pronounced spatial
461 component may thus have tapped into the tamarins' adaptive behaviour as extractive foragers
462 (Peres, 1992) and thereby made the tasks easier to solve. However, whenever the aim is to
463 assess an individual's cognitive ability per se, rather than a specific adaptation to a narrow
464 cognitive problem, it is preferable to present a task in an abstract rather than a naturalistic
465 manner. This is perhaps most evident for general intelligence test batteries that consist of a
466 number of subtests assessing a wide range of abilities from various cognitive domains to
467 identify whether they are all based on a single domain-general cognitive ability (Burkart,
468 Schubiger & van Schaik, *in press*). The traditional 2-choice memory test in experiment 1 is a
469 subtest of one such test battery, and we developed our 9-choice memory test as a possible
470 alternative.

471

472 **General discussion**

473 We conducted two experiments in order to explore whether the task format affects
474 cognitive performance of non-human primates in physical cognition tests just as it has been
475 reported for a widely used social cognition test (Burkart & Heschl, 2006). When testing
476 marmosets with a traditional 2-choice memory test (experiment 1), we found that, in contrast to
477 the Ebbinghaus effect, their performance did not continuously decline with increasing delay

478 duration. To address the possibility that our results reflected a lack of motivation to actually
479 memorize the location of the food rather than the marmosets' ability to do so, we designed a
480 new version of the memory test (experiment 2) with nine choice options instead of two, which
481 lowered the probability of making a correct choice by chance from 50% to 11%. Both
482 marmosets and squirrel monkeys performed better in the 9-choice memory test, and their
483 performance now continuously decreased with increasing delay duration, consistent with the
484 Ebbinghaus effect we had predicted. Our results suggest that the 9-choice format is more
485 accurate in assessing memory performance in the two New World monkey species, and that the
486 2-choice format negatively affects performance not only in a social cognition task, but also in a
487 physical one.

488 Our findings have important implications for studies that assess cognitive performance
489 in non-human primates and other animals for comparative purposes. Examples of such
490 comparisons include the assessment of differences in cognitive performance across different
491 tasks between individuals of one species (e.g., to investigate general intelligence; Banerjee et
492 al., 2009; Herrmann, Hernández-Lloreda, Call, Hare & Tomasello, 2010), between conspecifics
493 differing in certain traits (e.g., to investigate sex differences; Schubiger et al., 2015) or
494 environmental/ontogenetic conditions (e.g., to investigate rearing differences; Damerius &
495 Forss et al. *in prep.*; Hermann & Call, 2012), and differences in cognitive performance between
496 species (i.e., to investigate evolutionary trajectories; Amici, Aureli & Call, 2008, 2010). For all
497 these comparative purposes it is crucial that differences in measured performance reflect true
498 differences in the subjects' cognitive abilities and cannot be attributed to differences in their
499 motivation to engage with a specific task.

500 Decreasing the chance-level probability of success, as we have done in the present study,
501 is one way of promoting the subjects' motivation. But although using more than two choice
502 options is advantageous in some cognitive tests with animals, it is probably not feasible in

503 others. Examples for physical cognition tests that require the 2-choice format are the ones in
504 which the subject has to base its choice on more or less apparent differences in the amount (e.g.
505 numerical discrimination tests, Agrillo, 2014), or external features (e.g. tool functionality,
506 Mulcahy & Schubiger, 2014) of the test stimuli. In such tests, additional options could either
507 lead to ambiguous choices or be too demanding for a subject's working memory. However, the
508 costs of a wrong choice can also be increased in 2-choice tests, e.g. by requiring subjects to
509 choose by performing an effortful behavioural response such unscrewing a lid, pulling in the
510 chosen item, or a similarly effortful behaviour.

511 In sum, we found that non-human primates may not be sufficiently motivated to fully
512 engage in a cognitive task when presented in a 2-choice format but that some methodological
513 modifications can restore their motivation.

514 If future studies show that our findings generalize to other species beyond marmosets
515 and squirrel monkeys, and to cognitive domains other than memory, it may be preferable to
516 replace the 2-choice format with alternative task formats. Otherwise, cognitive performance
517 may be biased both in comparisons within and across species, for instance toward more food
518 motivated individuals or species.

519

520 **Acknowledgements**

521 We thank Carel van Schaik and Nick Mulcahy for discussion, Erik Willems for statistical
522 advice, head animal keeper, Heinz Galli, for help with building the experimental apparatuses,
523 and animal keepers, Thomas Bischof and Patricia Rivera, for taking good care of the monkeys.
524 We also thank two anonymous reviewers for their helpful comments. This study was
525 financially supported by the Swiss National Science Foundation (project number 310030-
526 130383) and the A. H. Schultz-Foundation. It was conducted under guidelines established by

527 the National Veterinary Office of Switzerland and licensed to be conducted by the Veterinary
528 Office of the Canton of Zurich (license number 183/13).

529

530 **References**

531

532 Agrillo, C. (2014). Numerical and arithmetic abilities in non-primate species. In Kadosh, R. C.,
533 & Dowker, A. (Eds.), .The *Oxford Handbook of Numerical Cognition* (Chapter 12). Oxford
534 University Press: Oxford, UK. DOI:10.1093/oxfordhb/9780199642342.001.0001

535

536 Amici, F., Aureli, F., & Call, J. (2008). Fission-fusion dynamics, behavioral flexibility, and
537 inhibitory control in primates. *Current Biology*, 18(18), 1415-1419.
538 DOI:10.1016/j.cub.2008.08.020

539

540 Amici, F., Aureli, F., & Call, J. (2010). Monkeys and apes: are their cognitive skills really so
541 different? *American Journal of Physical Anthropology*, 143(2), 188-197.
542 DOI:10.1002/ajpa.21305

543

544 Banerjee, K., Chabris, C. F., Johnson, V. E., Lee, J. J., Tsao, F., & Hauser, M. D. (2009).
545 General intelligence in another primate: individual differences across cognitive task
546 performance in a New World monkey (*Saguinus oedipus*). *PLoS One*, 4(6), e5883.
547 doi:10.1371/journal.pone.0005883

548

549 Burkart, J., & Heschl, A. (2006). Geometrical gaze following in common marmosets (*Callithrix*
550 *jacchus*). *Journal of Comparative Psychology*, 120(2), 120-130. DOI:10.1037/0735-
551 7036.120.2.120

552

553 Burkart, J. M., Schubiger, M. N., & van Schaik, C. P. The evolution of general intelligence.
554 Behavioral and Brain Sciences. BBS-D-16-00117R1.
555
556 Call, J. (2006). Inferences by exclusion in the great apes: the effect of age and species. *Animal*
557 *Cognition*, 9(4), 393-403. DOI:10.1007/s10071-006-0037-4
558
559 Deaner, R. O., Van Schaik, C. P., & Johnson, V. E. (2006). Do some taxa have better domain-
560 general cognition than others? A meta-analysis of nonhuman primate studies. *Evolutionary*
561 *Psychology*, 4(1), 149-196. DOI: 10.1177/147470490600400114
562
563 Ebbinghaus, H. (1885). *Über das Gedächtnis. Untersuchungen zur experimentellen*
564 *Psychologie*. Duncker & Humber, Leipzig.
565
566 Ebbinghaus, H. (1913). *Memory: A contribution to experimental psychology* (H. A. Ruger & C.
567 E. Bussenius, Trans.). New York: Columbia University, Teachers College. (Original work
568 published 1885).
569
570 Erdőhegyi À., Topál J., Virányi Z., & Miklósi Á. (2007). Dog-logic: inferential reasoning in a
571 two-way choice task and its restricted use. *Animal Behaviour* 74, 725–737. DOI:
572 10.1016/j.anbehav.2007.03.004
573
574 Fischer, G. J., & Kitchener, S. L. (1965). Comparative learning in young gorillas and
575 orangutans. *Journal of Genetic Psychology*, 107(2), 337-348.
576
577 French, G. M. (1959). Performance of squirrel monkeys on variants of delayed response.
578 *Journal of Comparative and Physiological Psychology*, 52, 741-745.

579

580 Harlow H. F. (1932). Comparative behavior of primates. III. Complicated delayed reaction tests
581 on primates. *Journal of Comparative Psychology*, 14:241–252.

582

583 Herrmann, E., Hernández-Lloreda, M. V., Call, C., Hare, B., & Tomasello, M. (2010). The
584 structure of individual differences in the cognitive abilities of children and chimpanzees.
585 *Psychological Science*, 21(1), 102-110.

586

587 Herrmann, E., & Call, J. (2012). Are there geniuses among the apes? *Philosophical*
588 *Transactions of the Royal Society of London. Series B, Biological Sciences*, 367(1603), 2753-
589 2761.

590

591 Hurvich, C. M., Tsai, C.-L. (1989). Regression and time series model selection in small
592 samples, *Biometrika*, 76, 297–307.

593

594 Kendrick, D. F., Rilling, M. E., & Denny, M. R. (1986). *Theories of animal memory*. Hillsdale,
595 N.J: L. Erlbaum Associates.

596

597 Lind, J., Enquist, M., & Ghirlanda, S. (2015). Animal memory: A review of delayed matching-
598 to-sample data. *Behavioural processes*, 117, 52-58.

599

600 Menzel, E., Juno, C., & Garrud, P. (1985). Social foraging in marmoset monkeys and the
601 question of intelligence [and discussion]. *Philosophical Transactions of the Royal Society B:*
602 *Biological Sciences*, 308(1135), 145-158.

603

604 Mikolasch, S., Kotrschal, K., & Schloegl, C. (2011). Is caching the key to exclusion in

605 corvids? The case of carrion crows (*Corvus corone corone*). *Animal Cognition*, 15, 73–82. DOI
606 10.1007/s10071-011-0434-1

607

608 Miles, R. C. (1957a). Delayed-response learning in the marmoset and the macaque. *Journal of*
609 *Comparative and Physiological Psychology*, 50, 352-355.

610

611 Miles, R. C. (1957b). Learning-set formation in the squirrel monkey. *Journal of Comparative*
612 *and Physiological Psychology*, 50(4), 356-357.

613

614 Miles, R. C., & Meyer, D. R. (1956). Learning sets in marmosets. *Journal of Comparative and*
615 *Physiological Psychology*, 49, 219-222.

616

617 Mulcahy, N. J., & Hedge, V. (2012). Are great apes tested with an object object-choice task?
618 *Animal Behaviour*, 83(2), 313-321. DOI:10.1016/j.anbehav.2011.11.019

619

620 Mulcahy, N. J., & Schubiger, M. N. (2014). Can orangutans (*Pongo abelii*) infer tool
621 functionality? *Animal Cognition*, 17(3), 657-669. DOI:10.1007/s10071-013-0697-9

622

623 Peres, C. A. (1992). Prey-capture benefits in a mixed-species group of Amazonian tamarins,
624 *Saguinus fuscicollis* and *S. mystax*. *Behavioral Ecology and Sociobiology*, 31, 339–347. DOI:
625 10.1007/BF00177774.

626

627 Petter, M, Musolino, E., Roberts, W. A. & Cole M. (2009). Can dogs (*Canis familiaris*) detect
628 human deception? *Behavioural Processes*, 82, 109–118. DOI:10.1016/j.beproc.2009.07.002

629

630 Reader, S. M., Hager, Y., & Laland, K. N. (2011). The evolution of primate general and
631 cultural intelligence. *Philosophical Transactions of the Royal Society B: Biological Sciences*,
632 366(1567), 1017-1027. DOI:10.1098/rstb.2010.0342
633

634 Rodriguez J. S., Paule M. G. (2009). Working Memory Delayed Response Tasks in Monkeys.
635 In Buccafusco J. J, (Ed.). *Methods of Behavior Analysis in Neuroscience* (2 nd ed.). Boca
636 Raton (FL): CRC Press/Taylor & Francis.
637

638 Rowe, N. (1996). *The pictorial guide to the living primates*. New York: Pogonias Press.
639

640 Schubiger, M. N., Wustholz, F. L., Wunder, A., & Burkart, J. M. (2015). High emotional
641 reactivity toward an experimenter affects participation, but not performance, in cognitive tests
642 with common marmosets (*Callithrix jacchus*). *Animal Cognition*, 18(3), 701-712.
643 doi:10.1007/s10071-015-0837-5
644

645 Tomasello, M., & Call, J. (1997). *Primate Cognition*. New York: Oxford University Press.
646

647 Treichler, F (1964). Delayed-response performance by the squirrel monkey. *Psychonomic*
648 *Science* 1(1), 129-130. DOI:10.3758/BF03342825
649

650 Tsujimoto, S., & Sawaguchi, T. (2002). Working memory of action: a comparative study of
651 ability to selecting response based on previous action in New World monkeys (*Saimiri sciureus*
652 and *Callithrix jacchus*). *Behavioural Processes*, 58(3), 149-155.
653

654 Woodruff, G., & Premack, D. (1979). Intentional communication in the chimpanzee: The
655 development of deception. *Cognition*, 7, 333-362.