Supplementary Information for

Chimpanzees flexibly update working memory contents and show susceptibility to distraction in the self-ordered search task

Christoph J. Völter, Roger Mundry, Josep Call, Amanda M. Seed

Corresponding author: Christoph J. Völter
Email: christoph.voelter@vetmeduni.ac.at

This supplementary file includes:

Supplementary text
Figs. S1 to S5
Tables S1 to S6
Captions for movies S1 to S2

Other supplementary materials for this manuscript include the following:

Movies S1 to S2
Datasets S1

Supplementary text

Experiment 1: Initial assessment

Setup. The experimenter sat behind a sliding platform (78.5 x 34 cm) facing the subject who was behind a mesh panel. Markings on the platforms subdivided it into seven (in the re-test phase 10) equally sized sections. In the initial assessment phase, two to six opaque boxes with lids served as hiding places of food rewards (half banana pellets). In all three conditions (Feature+Space, Space-Only, Feature-Only), we increased the number of boxes in a step-wise manner from two to six (re-test phase: from four to ten boxes) depending on subjects' performance. In the Feature+Space and Space-Only conditions, the location of each box on the platform was constant across subjects and trials. There were seven positions on the sliding table. We started with two boxes at the two innermost positions (3 and 4, numbered from left to right) and added boxes to the left and the right of these boxes depending on whether individuals reached a predetermined criterion with a given number of boxes. In the Feature-Only condition, we used the same positions on the platform, but the boxes were transferred to a second, adjacent (but otherwise identical) platform after each choice.

GLMM fitting and assumptions. We examined variables that predicted whether subjects committed an error or not in the Feature+Space condition by coding every *opportunity* for committing an error separately. That is, for every choice within a trial, we coded every empty (i.e., previously visited) box separately and scored whether or not apes chose the empty boxes again. We used a Generalized Linear Mixed Model [GLMM 01; 1] with binomial error structure and logit link function [2] to analyse these data. We examined the cues potentially used by chimpanzees to remember their choices by exploring the effect of condition in a second GLMM (GLMM 02) with binomial error structure and logit link function. In both models subject ID and trial ID (and choice ID in GLMM 01; with trial ID and choice ID being nested within subject) were included as random effects. To keep type I error rate at the nominal level of 5%, we included all possible random slope components (except for the correlation parameters among random intercepts and random slopes terms) [3, 4]. In GLMM 01, we examined the effect of:

the distance between revisits (i.e., number of visits between revisits within the same trial), whether subjects had made any mistake within this trial before, the spatial position of the boxes on the platform (inner boxes vs. outer boxes), the number of boxes on the platform (3-6, the two-box trials were excluded from this analysis because there were no inner boxes), the time interval (in seconds) subjects were absent from the platform during the retention interval before each choice, the response latency (in seconds) to make a choice (starting when the platform was pushed forward), sex, and age as test predictors and the trial number as control predictor. We included an offset term to control for varying probabilities for mistakenly choosing a particular empty box (log(1/number of empty boxes)) [2]. Moreover, subject ID, choice ID, and trial ID were included as random effects. To keep type I error rate at the nominal level of 5% [3, 4], we included all random slope components (except for the correlation parameters among random intercepts and random slopes terms) of distance between revisits, any error earlier, number of boxes, outer box, absence interval, response latency, and trial number within subject ID and outer box also within trial ID.

We examined the cues used by chimpanzees to remember their choices by exploring the effect of condition in a second GLMM (GLMM 02) with binomial error structure and logit link function. This analyzed whether apes emptied all boxes in a trial without any redundant search or not. We included condition (Feature + Space, Feature Only, Space Only), number of boxes (2-6), and their interaction as well as sex and age as test predictors and trial (within each condition and number of boxes) and session number as control predictors. We included these predictors as fixed effects and subject ID and trial ID (nested within subject ID) as random effect. We included random slopes components of condition (manually coded and then centered), number of boxes and their interaction as well as trial number and session number within subject ID except for the correlation parameters between random intercepts and random slopes terms.

Prior to fitting the models, the covariates were z-transformed (to a mean of zero and a standard deviation of one) to make the estimates easier to interpret. We determined variance inflation factors [5] for standard linear models excluding the random effects using the R package car [6]. Collinearity was no issue in GLMM 1 (maximum Variance Inflation Factor (VIF): 2.3 for age). In GLMM 2, VIF revealed some degree of collinearity between condition and session (VIF: 5.5 and 5.6, respectively), which was unsurprising given that the feature + space condition was always administered first. We assessed model stability by comparing the estimates derived from the model based on all data with those obtained from models with individual subjects and trials and also choice ID in model 1 (i.e., the levels of the random effects) excluded one at a time. This revealed the models to be stable with regard to the fixed effects.

As an overall test of the effect of the test predictors we compared each full model with a respective null model lacking the test predictors but comprising the same control predictors, offset terms, and random effects structure as the full model [7] using a likelihood ratio test [8]. P values for the individual effects were based on likelihood ratio tests comparing the full with respective reduced models [3; R function drop1 with argument 'test' set to "Chisq"]. The p values for the post-hoc pairwise comparisons of factor levels were adjusted for multiple comparisons [using the single-step method of the glht function, R package multcomp; see 9]. The model was implemented in R [version 3.3.2; 10] using the function glmer of the R package lme4 [11]. Confidence intervals for the binomial models were derived using the function bootMer of the R package lme4, using 1,000 parametric bootstraps and bootstrapping over the random effects.

The sample of GLMM 1 consisted of 1479 opportunities to choose an empty box of 9 chimpanzees who performed 702 choices within 250 trials. GLMM 2 included 719 trials of 9 chimpanzees. The data are available as part of the supplementary material.

Search strategies. For the assessment of potential search strategies, we calculated two different indices for trials with more than 2 boxes: a linear search index and a serial ordering index. For the linear search index, we scored 1 for each choice within a trial (excluding the first choice) if the preceding choice was an adjacent box and 0 if there was at least one box in between. For every trial we then calculated the mean of these linear search scores. For the serial ordering index, we scored 1 for each choice within a trial if apes' search order complied with the order in which they were presented with the boxes throughout the experiment. When we increased the number of boxes within the experiment we added the same boxes at the same location for all subjects in the feature + space condition. For every trial, we then calculated the mean of these serial ordering scores. This measure therefore captures the degree of familiarity with the different boxes and locations. To analyse whether apes' performance benefited from these search strategies we calculated Spearman correlations between the search indices (linear search and serial ordering) and accuracy per individual and condition (feature / space).

We found a significant correlation between the linear search index and accuracy only in the space-only condition ($r_S = 0.765$, N = 9, *p* = 0.021, see Fig.S1) but not in the feature + space condition ($r_S = 0.017$, N = 9, *p* = 0.981) or the feature-only condition ($r_S = 0.486$, N = 6, *p* = 0.356). In contrast, we found no correlation between the serial ordering index and accuracy in any of the conditions (space-only: $r_S = 0.477$, N = 9, *p* = 0.198; feature + space: $r_S = -0.477$, N = 9, *p* = 0.200; feature-only: $r_S = 0.029$, N = 6, *p* = 1).

Experiment 1: retest phase

Subjects. We tested the same nine chimpanzees 9 to 10 months after they had completed the feature + space condition of the initial assessment phase of Experiment 1 (and ca. 8 months after they had completed Experiment 2).

Materials, procedure, and design. We used the same setup and procedure as in the initial assessment phase a with few modifications. First, we reduced the spacing on the sliding platform to accommodate 8 boxes. We used the same 6 boxes as in the feature + space condition of the initial assessment phase and added two more boxes for the individuals that passed the 6 and 7 boxes conditions. For one individual (Kofi) who passed the 8 box condition we used an entirely new set of slightly smaller boxes after he completed the 8-box condition. For this individual, we used a modified sliding platform for the new set of boxes with a narrower spacing to accommodate ten boxes.

All individuals started with the 4-box stage. Depending on their performance we increased the number of boxes until they reached 10 boxes. As we had to change the boxes when a subject completed the 8-box condition we repeated the 8-box condition with the novel set to ensure that the novel boxes did not affect chimpanzees' performance.

We used the same test criterion as in the initial assessment phase to decide whether an individual would receive the next higher number of boxes (two consecutive trials correct). However, we did not stop data collection until subjects got three consecutive trials correct (or until they reached the maximum trial number; same as in the initial assessment phase, 8 trials for the 7 to 10-box condition) to get a more sensitive measure of their memory capacity at the individual level.

Scoring and analysis. Regarding potential search strategies, we examined in addition to linear search strategies whether the variability of chimpanzees' search behaviour was related to their accuracy. One could hypothesize that the successful individuals have idiosyncratic but highly conservative search strategies across trials [12]. Such a strategy could reduce the short-term memory load considerably. To investigate whether some individuals might have acquired such idiosyncratic search strategies we calculated a search variability index. For every individual and number of boxes we counted, separately for each position in the search sequence, the number of unique boxes chosen across the first three trials per number of boxes (the minimal number of trials completed by every individual). For example, if an individual across the three trials always chose the same box first we would assign "1" if the individual chose three times a different box as first choice, we assigned "3". Across all choices within the search sequence of a given number of boxes we calculated a mean score for every individual and number of boxes. We calculated correlations between search variability and accuracy per individual and number of boxes (we report only the correlations for 4 and 5 boxes because the sample size declined to 4 individuals with 6 boxes).

We used a GLMM (GLMM S01) with binomial error structure and logit link function to analyse all opportunities to choose an empty box in the retest phase. We only included the data up to 8 boxes (excluding all trials with the novel set of boxes) because there was only one individual left who passed the 8-box condition. The model was fitted in the same manner as GLMM 01. The only exception was that we did not include 'sex' as control variable due to convergence issues. We dropped 'sex' as control variable because it did not appear to have a noticeable effect on the error rates in the initial assessment phase. Collinearity was no issue in GLMM S01 (maximum VIF: 1.37 for Number of boxes). The sample of GLMM S01 consisted of 1979 opportunities to choose an empty box for 9 chimpanzees who performed 739 choices within 177 trials.

Finally, we calculated Spearman correlations to assess test-retest reliability. We used the mean individual performance with 4 and 5 boxes because all 9 individuals completed these conditions in the retest phase. Moreover, we ordered individuals according to the maximum number of boxes in which they reached the criterion. When individuals reached the same maximal number of boxes, we ordered them according to the number of trials they needed to reach the criterion in this condition. Based on this ranking we calculated Spearman correlations to examine the stability of individual performance limits across the two assessment phases.

Results: Search strategies and error rates. Similar to the initial assessment, we found no significant correlation between the linear search index and accuracy ($r_S = -0.100$, N = 9, $p = 0.811$). In addition, we found no evidence that variability in their search behaviour across trials was correlated with accuracy in the 4-box condition ($r_S = -0.185$, N = 9, $p = 0.684$) or in the 5-box condition ($r_S = 0.583$, N = 9, $p = 0.107$). In the first three trials individuals tended to visit on average 2 (with 4 boxes; range: 1.5 to 2.25) and 2.2 (with 5 boxes; range 1.8 to 2.6) unique boxes at any position in the search sequence indicating considerable variability in search patterns across trials (the values of the variability index could range between 1 and 3).

We fitted a GLMM (GLMM S1) to identify predictors of the subjects' probability to revisit an empty box. Similar to the initial assessment phase, we analysed for every box on the platform that the apes had chosen before (within the same trial) the probability that apes would revisit the box and included the predictors distancetime lag between revisits, whether they had made any mistake within this trial before, the number of boxes (4 to 8) on the platform, the spatial position of the boxes on the platform (inner boxes vs. outer boxes), age, and the trial number.

The full model fitted the data significantly better than a null model lacking the test predictors ($\chi^2(5) = 46.86$, $p < 0.001$; see ESM Table S3 for detailed results). We found that the longer the distance between revisits, the higher was the apes' probability to revisit the box ($\chi^2(1) =$

13.05, $p < 0.001$; see Fig. S2*a*). Moreover, chimpanzees were less likely to revisit outer boxes (compared to inner boxes; $\chi^2(1) = 22.83$, $p < 0.001$; see Fig S2*b*). Finally younger apes made less mistakes than older ones ($\chi^2(1) = 10.58$, $p = 0.001$). The number of boxes did not affect the error probability per empty box ($\chi^2(1) = 1.03$, $p = 0.309$). Whether apes had made a mistake within the same trial before or not did not significantly affect the probability to make a mistake in the current choice ($\chi^2(1) = 0.81$, $p = 0.369$). Trial number ($\chi^2(1) = 0.06$, $p = 0.813$) did not have obvious effects on error rates either.

Results: Memory capacity. In the re-test phase, we used a stricter test criterion of three consecutive trials correct and added more boxes to the search array, which allowed us to compare chimpanzees' individual performance to simulations of different memory sizes. One chimpanzee (Kofi), performed significantly better than a memory size (MS) 7 simulation (with 10 boxes on the platform). Sandra performed better than the MS 4 simulation with 7 boxes, Lome performed better than MS 2 simulation with 5 boxes, four chimpanzees performed better than the MS 1 simulation with 4 boxes, and two individuals performed better than chance with four boxes (all $p < 0.05$).

Given that Kofi's performance surpassed all of the other chimpanzees we examined his search behaviour in more detail. Table S3 shows Kofi's performance in his final trials with 8 to 10 boxes. His search pattern did not appear to be completely random but also not linear. Most notable was his tendency to finish his search with the outer boxes (that are associated with lower error rates, see GLMM S01). However, his search pattern was not constant to an extent that would reduce the memory load in any obvious way. For example, across the 8 trials with 10 boxes Kofi chose on average 5 different boxes at each point in his search sequence (range: 3 to 6). The search pattern (i.e., when in the sequence he visited which box) of his last two successful trials with 10 boxes did not overlap at all.

Experiment 2

GLMM details

The model was stable with regard to the effects of condition, trial number, and order of conditions but rather unstable for sex and age when subjects were excluded one at a time. Collinearity (maximum VIF: 3.91 for age and sex) appeared to be no issue. The data for GLMM 3 included 144 trials of 8 individuals. The data are available as part of the supplementary material.

Platform 1. We compared individual performance in the Identical Boxes and Different Boxes conditions to chimpanzees' performance in Experiment 1 (initial assessment: feature + space condition) and its retest phase. We found that apes' performance in the Different Boxes condition was correlated with their performance in feature + space condition (Experiment 1 - initial assessment: $r_S = 0.717$, N = 9, $p = 0.035$; retest: $r_S = 0.741$, N = 9, $p = 0.028$). In the Identical Boxes condition the pattern of correlations was more mixed (initial assessment: $r_S = 0.364$, N = 9, $p = 0.331$; retest: $r_S = 0.756$, N = 9, $p = 0.026$).

Platform 2 (GLMM S02). In GLMM S02, we analysed chimpanzees' platform 2 performance of the Different Boxes and Identical Boxes conditions. The Food Distraction condition was not included here because there were no boxes on platform 2 in the Food Distraction condition. Apart from this, the model specification was identical to GLMM 03.The data used for GLMM S02 consisted of 96 trials of 8 chimpanzees.

GLMM S02, comprising the test predictors condition (DB or IB), and age, along with the control predictors order of condition and trial number, fitted the data significantly better than null model comprising only the control predictors and the random effects ($\chi^2(3) = 10.80$, $p = 0.013$, see Table S6 for detailed results). There was no significant difference between the Identical Boxes and the Different Boxes condition at platform 2 ($\chi^2(1) = 3.12$, $p = 0.078$) but a trend toward better performance in the Different Boxes than Identical Boxes condition. Younger subjects performed better than older ones ($\chi^2(1) = 8.14$, $p = 0.004$) whereas sex ($\chi^2(1) = 3.59$, $p = 0.058$) did not have a significant effect on performance. The control predictors order of condition and trial number had no significant effects on performance (both $p > 0.1$).

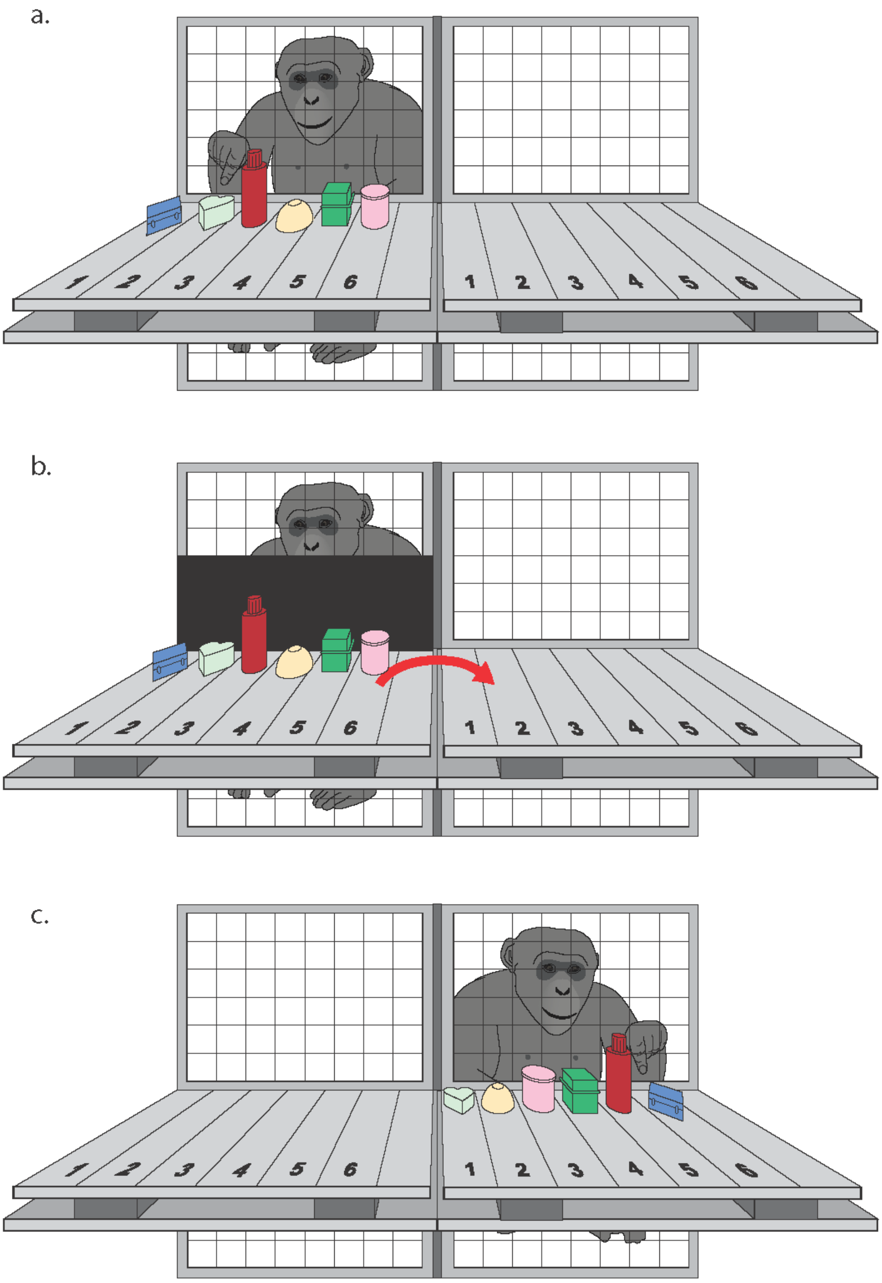

Fig. S1 Illustration of the setup and procedure of the Feature Only condition in Experiment 1: *a*: starting position on platform 1 with 6 boxes; *b*: after the first choice the experimenter (E) occludes platform 1 and transfers all the boxes to the adjacent platform 2 (red arrow); thereby, E changes the order of boxes; *c*: the subject makes the second choice. After the second choice, the experimenter transfers the boxes back to platform 1 (not depicted). This procedure is repeated until all the boxes could have been emptied without redundant choices.

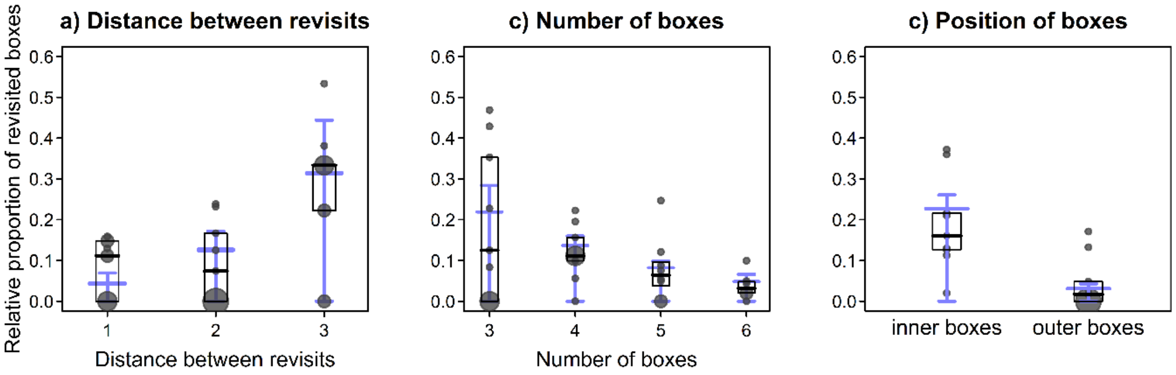

Fig. S2 Experiment 1 (GLMM 01): Box plots of factors that predicted the apes' probability to commit an error (i.e., revisiting a box). The number of revisited boxes divided by the relative frequency of empty boxes on the platform is plotted (means per individual) as a function of *a*: distance between revisits (1-3; the 4-boxes condition serves here to visualize the effect), *b*: the number of boxes on the platform, and *c*: the position of the boxes on the platform (outer vs. inner position in the array of boxes). The boxes indicate the quartiles and the horizontal lines inside the boxes show median values. The blue vertical lines depict the bootstrapped 95% confidence intervals of the fitted model, the blue (wide) horizontal lines depict the model estimates. The area of the dots corresponds to the number of individuals per condition and relative proportion of revisited boxes (N = 1 to 5).

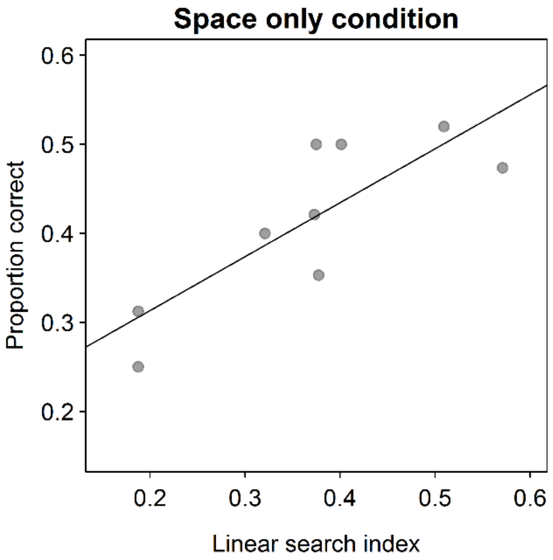

Fig S3. Experiment 1: Mean proportion of correct trials in the space only condition plotted against the linear search strategy.

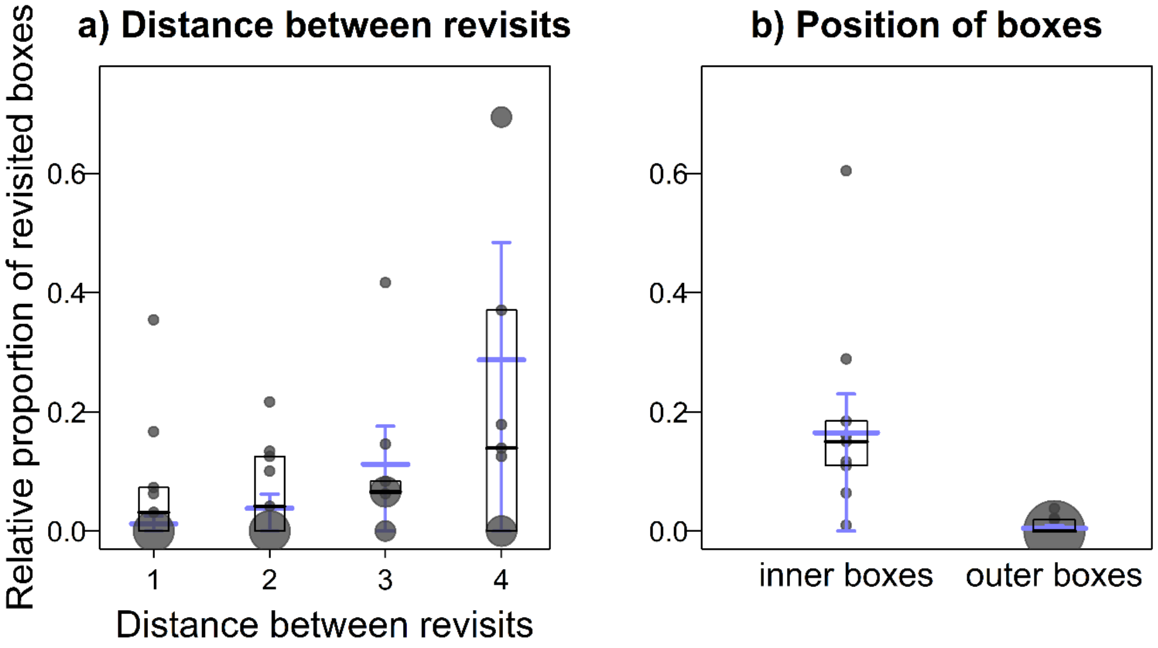

Fig. S4. Retest phase of Experiment 1: Box plots of factors that predicted the apes' probability to commit an error (i.e., choosing an already emptied box). The proportion of revisited empty boxes divided by the relative frequency of empty boxes on the platform is depicted (means per individual) as a function of a) the distance between revisits (1 to 4; the 5-boxes condition serves here to visualise the effect) and b) the position of the boxes on the platform (outer vs. inner position in the array of boxes). The boxes indicate the quartiles and the black horizontal lines inside the boxes show median values. The blue vertical lines depict the bootstrapped 95% confidence intervals of the fitted model, the blue (wide) horizontal lines depict the model estimates. The area of the dots depicts the number of individuals per condition and relative proportion of revisited boxes (N = 1 to 7).

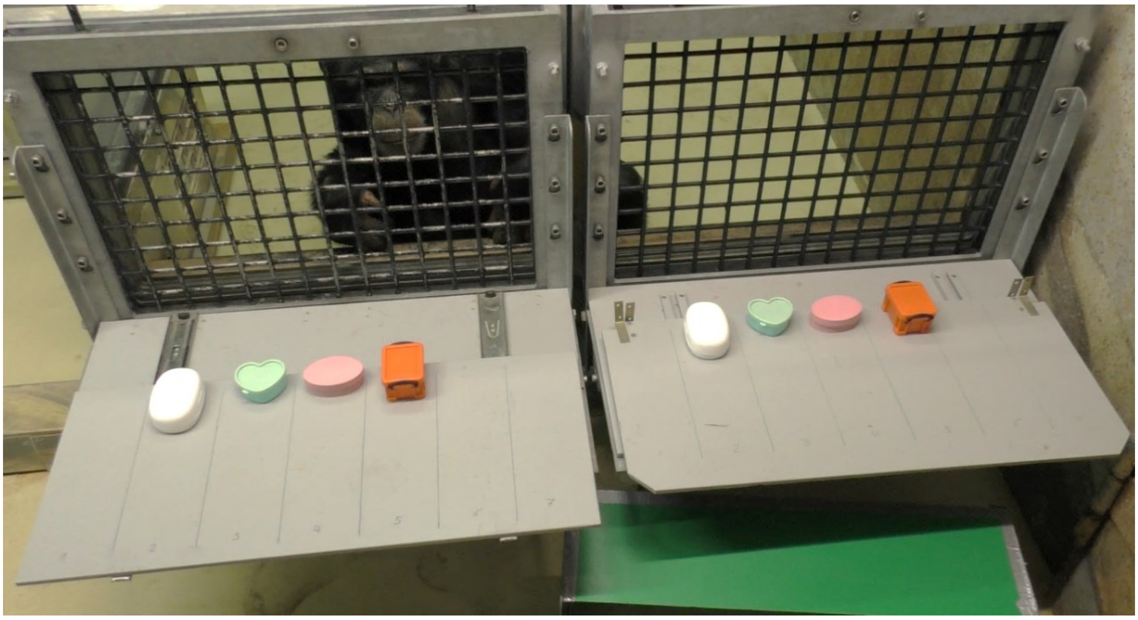

Fig. S5. Setup of the identical boxes condition in Experiment 2.

Table S1. Results of GLMM 01: Analysis of error rates in the initial feature+space condition of Experiment 1

	Estimate	SE	X²	DF	P	95% CI	
(Intercept)	-1.311	0.242				-11.813	-1.031
Distance between revisits[4]	0.671	0.137	11.951	1	0.001	0.410	2.493
Any error earlier[1]	0.511	0.430	0.965	1	0.326	-0.754	2.795
Number of boxes[5]	-0.545	0.128	13.345	1	<0.001	-2.240	-0.351
Position of boxes[2]	-2.200	0.304	17.855	1	<0.001	-7.943	-1.803
Absence interval[5]	0.389	0.112	6.612	1	0.010	0.155	1.005
Response latency[6]	0.040	0.079	0.259	1	0.611	-0.358	0.534
Sex[3]	0.132	0.361	0.166	1	0.684	-0.718	1.000
Age[6]	0.447	0.166	7.992	1	0.005	0.128	1.145
Trial number[7]	-0.259	0.118	3.792	1	0.051	-0.727	-0.027

Notes: Reference categories: [1]no error earlier, [2]inner boxes, [3]female. Covariates were z-transformed to a mean of zero and a standard deviation of one; mean (sd) of the original variable were [4]1.94 (1.04), [5]2.27 (6.84), [6]0.32 (1.57), [7]4.87 (0.97), [8]27.27 (11.40), [9]4.16 (2.60).

Table S2. Results of GLMM 02: Correct choices across the different conditions of Experiment 1

	Estimate	SE	X²	DF	P	95% CI	
(Intercept)	2.166	0.508				1.319	3.202
Condition[1]: feature only	-2.845	0.821				-4.494	-1.345
Condition[1]: space only	-1.591	0.811				-3.187	-0.102
Number of boxes[3]	-1.188	0.295				-1.798	-0.642
Sex[2]	-0.721	0.405	3.002	1	0.083	-1.482	0.037
Age[4]	-0.485	0.202	5.215	1	0.022	-0.912	-0.090
Session number[5]	0.831	0.404	4.112	1	0.043	0.046	1.718
Trial number[6]	0.084	0.107	0.623	1	0.430	-0.131	0.343
Condition[1] x Number of boxes[3]			7.515	2	0.023		
Condition[1]: feature only x Number of boxes	-0.359	0.342				-1.074	0.269
Condition[1]: space only x Number of boxes	-1.319	0.464				-2.396	-0.561

Notes: Reference categories: [1]feature+space, [2]female. Covariates were z-transformed to a mean of zero and a standard deviation of one; mean (sd) of the original variable were [3]3.39 (1.31), [4]25.98 (11.32), [5]8.88 (5.11), [6]5.92 (4.85).

Table S3. Results of GLMM S1: Analysis of error rates in the retest phase of Experiment1

	Estimate	**SE**	**X^2**	**DF**	**P**	**95% CI**	
(Intercept)	-1.647	0.250				-2.509	-1.208
Distance between revisits[3]	1.382	0.261	13.054	1	0.000	0.904	1.995
Any error earlier[1]	0.342	0.370	0.808	1	0.369	-0.491	1.05
Number of boxes[4]	-0.225	0.212	1.034	1	0.309	-0.615	0.153
Position of boxes[2]	-3.810	0.523	22.828	1	0.000	-5.463	-3.074
Age[5]	0.949	0.236	10.583	1	0.001	0.523	1.487
Trial number[6]	0.027	0.113	0.056	1	0.813	-0.205	0.251

Notes: Reference categories: [1]no error earlier, [2]inner boxes. Covariates were z-transformed to a mean of zero and a standard deviation of one; mean (sd) of the original variable were [3]2.22 (1.28), [4]5.77 (1.33), [5]24.84 (10.63), [6]4.42 (2.68).

Table S4. Kofi’s spatial search pattern with 8 to 10 boxes on the platform. The position (I – X) shows the spatial distribution of the boxes on the sliding platform. The colour coding serves to highlight the search sequence (1 to 10); ‘-’ marks omission errors; ‘/’ marks commission errors (redundant searches).

	Position									
Number of boxes	**I**	**II**	**III**	**IV**	**V**	**VI**	**VII**	**VIII**	**IX**	**X**
8		8	4	3	2	7	6	5	1	
		8	6	4	3	5	7	2	1	
		8	4	2	3	5	7	6	1	
9	8	6	4	2	3	5	7	9	1	
	9	3/8	-	1	2	7	6	4	5	
	9	3/4	-	2	8	5	6	1	7	
	9	8	4	-	5	2/6	3	7	1	
	9	6	4	3	5	7	2	1	8	
	9	5	2	1	4	7	3	6	8	
	9	5	3	2	1	8	4	6	7	
10	10	5	7	3	4	9	6	1	8	2
	10	6	3	7	2	8	4	1	5	9
	10	9	6	2/8	7	4	3	1	5	-
	-	9	5	8	4/6	1	2	3	7	10
	-	9	8	5	2/3	7	4	1	6	10
	9	6	5	4	7	2	3	10	1	8
	8	9	3	5	1	6	2	4	7	10
	10	7	9	3	4	1	-	2/5	6	8

Table S5. Results of GLMM 03: correct choices on platform 1 in Experiment 2

	Estimate	SE	X^2	DF	P	95% CI	
(Intercept)	2.287	0.904				0.612	5.231
Condition[1]: Food distraction	-0.003	0.496				-1.111	1.037
Condition[1]: Identical boxes	-1.551	0.498				-2.870	-0.642
Order[3]	-0.029	0.203	0.020	1	0.888	-0.466	0.393
Trial[4]	-0.290	0.201	2.101	1	0.147	-0.771	0.109
Age[5]	-1.494	0.753	3.345	1	0.067	-3.835	-0.132
Sex[2]	-2.601	1.501	2.612	1	0.106	-7.258	0.301

Notes: Reference categories: [1]different boxes, [2]female. Covariates were z-transformed to a mean of zero and a standard deviation of one; mean (sd) of the original variable were [3]2.0 (0.82), [4] 3.50 (1.71), [5]24.50 (11.05).

Table S6. Results of GLMM S02: correct choices on platform 2 in Experiment 2

	Estimate	SE	Χ²	DF	P	95% CI	
(Intercept)	1.984	0.904				0.584	4.883
Condition[1]	-1.350	0.832	3.115	1	0.078	-3.257	0.025
Order[3]	-0.421	0.390	1.054	1	0.305	-1.518	0.247
Trial[4]	0.057	0.264	0.046	1	0.831	-0.461	0.606
Age[5]	-1.592	0.625	8.141	1	0.004	-3.506	-0.559
Sex[2]	-2.129	1.323	3.591	1	0.058	-6.047	-0.140

Notes: Reference categories: [1]different boxes, [2]female. Covariates were z-transformed to a mean of zero and a standard deviation of one; mean (sd) of the original variable were [3]2.0 (0.79), [4] 3.50 (1.72), [5]24.50 (11.07).

Supplementary references

1. Baayen R.H. 2008 *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge, UK, Cambridge University Press.
2. McCullagh P., Nelder J. 1989 *Generalized linear models*. London, UK, Chapman and Hall.
3. Barr D.J., Levy R., Scheepers C., Tily H.J. 2013 Random effects structure for confirmatory hypothesis testing: Keep it maximal. *J Mem Lang* **68**(3), 255-278.
4. Schielzeth H., Forstmeier W. 2009 Conclusions beyond support: overconfident estimates in mixed models. *Behav Ecol* **20**(2), 416-420.
5. Field A. 2005 *Discovering statistics with SPSS*. London, UK, Sage.
6. Fox J., Weisberg S. 2011 *An {R} Companion to Applied Regression*. 2nd ed. Thousand Oaks, California, Sage.
7. Forstmeier W., Schielzeth H. 2011 Cryptic multiple hypotheses testing in linear models: overestimated effect sizes and the winner's curse. *Behavioral Ecology and Sociobiology* **65**(1), 47-55.
8. Dobson A.J. 2002 *An introduction to generalized linear models*. Boca Raton, Chapman & Hall/CRC press.
9. Hothorn T., Bretz F., Westfall P. 2008 Simultaneous inference in general parametric models. *Biometrical journal* **50**(3), 346-363.
10. R Core Team. 2016 R: A language and environment for statistical computing. . (Vienna, Austria, R Foundation for Statistical Computing.
11. Bates D., Maechler M., Bolker B., Walker S. 2015 Fitting Linear Mixed-Effects Models Using lme4. *J Stat Softw* **67**, 1-48.
12. Owen A.M., Downes J.J., Sahakian B.J., Polkey C.E., Robbins T.W. 1990 Planning and spatial working memory following frontal lobe lesions in man. *Neuropsychologia* **28**(10), 1021-1034.

Captions for supplementary movies

Movie S1. Experiment 1 (retest phase): Kofi’s seventh trial with ten boxes in the Feature + Space condition is shown.

Movie S2. Experiment 1: Kofi’s first trial with six boxes in the Feature-Only condition is shown.