Supporting Information

Benson-Amram et al. 10.1073/pnas.1505913113

SI Text

Administration of Puzzle Box Trials in Zoos. We ran puzzle box trials with carnivores maintained at nine North American zoos: St. Louis Zoo, St. Louis; Bergen County Zoo, Paramus, NJ; Binder Park Zoo, Battle Creek, MI; Potter Park Zoo, Lansing, MI; Columbus Zoo, Columbus, OH; The Living Desert, Palm Desert, CA; Wild Canid Survival and Research Center, Eureka, MO; Turtle Back Zoo, West Orange, NJ; and Denver Zoo, Denver.

Both large and small puzzle boxes allowed subjects to see and smell the bait inside, and all worked in exactly the same way: the animal had to slide a simple bolt latch sideways to have the hinged door swing open so that the food inside could be accessed, or move the box around until it was oriented such that the bolt would fall open as described previously (24, 28) (Movie S1). Baits were chosen based on what zoo keepers told us was the favorite food of each individual tested. Baits ranged from bamboo to dead baby goats, but in all cases, the bait could not exit the box unless the door was opened.

We tested animals that ranged in size from roughly 2 to 300 kg, and therefore, we used two steel mesh puzzle boxes of different sizes. The larger box was 63.5 × 33 × 33 cm, and the smaller box was one-half that size. The smaller box was presented to species with an average body mass of <22 kg, such as river otters, kinkajous, sand cats, and other small-bodied carnivores (Dataset S1). The larger box was presented to species with an average body mass >22 kg, including snow leopards, wolves, bears, and other large-bodied species (Dataset S1). However, for cheetahs (species average body mass = 50 kg) and wild dogs (species average body mass = 22.05 kg), both the large and small boxes were used with some subjects (cheetahs: three individuals tested with the small box and six individuals tested with the large box; wild dogs: three individuals tested with the small box and two individuals tested with the large box). Boxes presented to these two species varied in size because of specific requests by animal keepers at three zoos for a smaller box size after the larger box had already been used for these species at other locations. This variation in box size used did not affect our results given that no individual cheetah opened any puzzle box, regardless of its size. Moreover, in wild dogs, individuals were roughly as likely to open the small (22% of trials: two of nine total trials) as the large box (16.7%: one of six total trials). Finally, we also examined whether a species' body mass influenced its ability to open the box within each of two box sizes in separate analyses from those described above. Using the same model shown in Table 2 but only for individuals given the large [effect of body mass: 95% credible interval (95% CI) = −4.93–14.32; pMCMC = 0.52] or small (effect of body mass: 95% CI = −8.55–4.4; pMCMC = 0.51) box, we found no effect of body mass on the ability to open the puzzle box.

Each subject was tested alone in its home enclosure, all subjects were fasted for 24 h before testing, and all trials were videotaped in their entirety. Each subject was briefly moved to an adjacent enclosure while the baited box was placed in the animal's home enclosure, the latch handle was set to protrude at a 90° angle from the door, and a tripod-mounted video camera was aimed to center on the box from just outside the home enclosure. The box was always oriented such that its door was at right angles to the focal plane of the camera. The subject was then moved back into its home enclosure for testing. All zoo testing was conducted by G.S. The experimenter either was hidden completely from the subject or filmed the trial from among the zoo patrons visiting the exhibit. Trials lasted 30 min or until the animal obtained the bait from the box (mean amount of time attempting to open the box = 329.2 s; minimum = 0 s; maximum = 1,683 s). The box was scrubbed with disinfectant (whichever one was routinely used at that facility) and rinsed between trials.

For most subjects, each trial started when the animal entered the test enclosure, because the puzzle box was immediately within its field of view. However, in a few cases, the enclosure was large and full of foliage, such that the animal could enter the test enclosure but remain unaware of the presence of the puzzle box. For these individuals, we started trials at the first instant when the subject had a direct line of sight to the puzzle box and its orientation behavior indicated that it had detected the box. Thus, in all cases, a trial started when an individual could see the puzzle box, and the period of exposure to the puzzle box was 30 min for all subjects. Given that no trial began until the subject was clearly aware of the presence of the puzzle box, any difference in trial start time caused by the size of the enclosure or the density of foliage is likely to be negligible in a 30-min time window.

We conducted 495 trials in total, because we tested one to nine individuals per each of 39 species (mean = 4.9 individuals; median = 5) (Table S1). At least three different individuals were tested in 23 of 39 species, but in 12 species, there were only two individuals tested, and in 4 species, only one individual was tested. The average number of trials conducted per individual (mean = 4.2 trials; median = 3; range = 1–10) varied among species (Table S1). Each individual within each species was tested with the puzzle box approximately three times, but if the subject did not succeed at opening the puzzle box by its third trial, we conducted no additional trials. Some individuals were tested in more than three trials, because when permitted by zoo staff to do so, we attempted to calculate learning curves for a different study.

Our results were not sensitive to variation in either the total number of individuals tested per species or the average number of trials conducted per individual. Specifically, we obtained the same results if we limit our analyses only to species in which at least three (Table S3) or four individuals (Table S4) were tested or only individuals to which we administered at least 3 separate trials (total number of trials per individual was 3–10) (Table S5) or when we restricted our analysis to examining only the first 3 trials for each individual (Table S6). These additional analyses indicated that our main conclusion was robust to different methods of analysis and different subsets of the data, and thus that brain size relative to body mass influenced the likelihood that carnivore species would open the puzzle box but that sociality and other variables did not.

As in all comparative studies that examine behavioral responses across a broad range of species, care must be taken to minimize variation among species because of the different ecologies of those species. In recent years, there also has been a lively debate over the possibility that cognitive tests often fail to account for variation in performance among subjects because of differences in motivation, body morphology, cue salience, or prior experiences (8, 44, 57, 58). In this study, we attempted to deal with these issues by restricting our experiments to species from a single order and controlling for factors, such as manual dexterity and neophobia (measured as the latency to approach the puzzle box). To bring each individual up to a high level of motivation before testing, we presented each animal with its favorite food as a reward for opening the puzzle box and fasted each animal for 24 h before trials. Thus, red pandas seemed just as highly motivated to solve the puzzle for a reward of bamboo as Amur tigers for a reward of raw meat. Additionally, we purposefully designed an artificial task that individuals would not likely have previously experienced (59, 60). This task did not require individuals to associate rewards with particular colors or shapes, which may introduce bias into results. Instead, the reward in this task was easily discernable from both visual and olfactory cues. In nine of the puzzle box trials, individuals opened the box door but did not retrieve the food reward, which might reflect underlying differences in motivation. We included these trials in our main analyses (Table 2) but also, ran our analyses without these nine trials and found the same qualitative results (Table S8). Interestingly, we found that species with a larger body masses were less successful than smaller species (results reported in the text). This finding may be an artifact of fasting all animals for the same 24-h time period before testing. It is possible that small species may be hungrier after a 24-h fasting period than larger species, and thus, motivation may be higher in smaller-bodied species. However, we think this possibility is unlikely given that the behavior of the animals did not indicate a lack of motivation to solve the problem. Latency to approach the puzzle box did not influence success in solving the problem (Table 2), and only one tested individual appeared uninterested in the food reward.

One factor that may have influenced performance on this task for which we could not fully account was the enrichment history of each captive subject. We tested carnivores housed in nine different zoos, and it is possible that individual animals experienced varying levels of enrichment and had different experiences with manmade objects before our tests. Although it would be ideal to test only individuals with identical prior experiences, it is difficult to imagine how large comparative tests where individuals from a range of species are presented with cognitive challenges could ever fully account for variation in enrichment histories. Even studies like that by MacLean et al. (12) compared results from species that were housed in many different research facilities around the world. It is quite likely that individuals in these facilities experienced varying exposures to manmade objects and even had different histories of participating in experiments where they were exposed to novel situations and cognitive challenges. One way to try and account for some of this variation is to include zoo identification as a random effect in our statistical models. This method assumes that individuals that are housed in the same zoo have similar enrichment experiences compared with individuals housed in different zoos. We did this and found that inclusion of zoo as a random effect had no effect on our main conclusion that carnivore species with larger brains for their body mass perform better in the puzzle box tests. Specifically, when zoo was included as a random effect in the top model shown in Table 1, where DIC was 283.2, the DIC value for this same model but now including zoo increased slightly to 283.5, suggesting that the addition of zoo as a random effect did not improve the overall fit of our top model. Furthermore, the effect of total brain volume on puzzle box performance in the top model shown in Table 2 but now including zoo as a random effect was similar in magnitude and direction (95% CI = 1.71–16.29; pMCMC = 0.016) as the top model shown in Table 1 without zoo as a random effect (Table 2). However, because not all species were tested in every zoo, we did not include zoo as a random effect in the models presented in our results, because including zoo leads to colinearity among the different random effects. Thus, although we have no reason to believe that the enrichment histories of the animals included in this study influenced our results, it would be ideal for future studies to try to minimize variation in enrichment histories as much as possible.

Extraction of Behavioral Data. S.B.-A. performed all data extraction from videotapes of zoo trials, and Adam Overstreet assisted with extracting measures of manual dexterity. We calculated interobserver reliability for our measure of dexterity. We had two observers extract dexterity measures for 10% of the trials, and our measure of interobserver reliability was very high (R = 0.94; Spearman rank correlation). Therefore, we are confident that our dexterity measures are highly reliable. We extracted performance measures from video footage as described earlier (24, 28). Briefly, we recorded the time taken to approach the puzzle box after the subject first detected the box (latency to approach box) as a measure of motivation to obtain the food reward. We scored work time as the number of seconds after trial onset during which the subject had its head down and was oriented toward and focused on the puzzle box before either opening the box or the trial ended. To score behavioral diversity, we looked for 13 behaviors of which all subjects were physically capable and scored whether an individual exhibited each behavior. Each individual, thus, received a score from 0 to 13. Here, we only scored each subject's first trial with the puzzle box. The 13 behaviors were rub, foot on box, sniff, lick, dig, bite, pull box with mouth, push box with head, push box with paw, pull box with paw, stand on box, tip box, and flip box. The highest score was 10 [achieved by three individuals: two coatis (*Nasua narica*) and one bobcat (*Lynx rufus*)], and the lowest score was 0 [nine individuals: one from each of the following species: bobcat (*L. rufus*), caracal (*Caracal caracal*), cheetah (*Acinonyx jubatus*), kinkajou (*Potos flavus*), red panda (*Ailurus fulgens*), ringtail (*Bassariscus astutus*), river otter (*Lontra canadensis*), sand cat (*Felis margarita*), and serval (*Leptailurus serval*)]. To score manual dexterity, we adopted methods from the work by Iwaniuk et al. (26) and scored each individual on 20 measures of forelimb movements, which we summed to calculate an overall measure of forelimb dexterity. The movements that we scored included

Body postures,

Limb crosses midline,

Alternate limb use,

Upper forelimb moves in more than one plane,

Upper forelimb rotation,

Lower forelimb rotation,

Grasping,

Picks up items,

Unimanual grasping,

Whole-forepaw grasp,

Digit 2–Digit 3 grasp,

Claw grasp,

Other grasp,

Independent digit movement,

Frequency of manipulation,

Items swapped between forepaws,

Items rotated by forepaw,

Distal digits used in manipulation,

Forepaws pull away from each other, and

Other forepaw movement.

We used no specific measure to quantify anxiety in our subjects during trials, but we did take all possible steps to minimize subjects' anxiety, such as having the videographer conceal himself from the subject after starting each video recording but before the subject was introduced back into its home enclosure. As can be seen in Movie S1 of our trials, our subjects generally appeared to be keenly interested in the food inside the box and did not exhibit behaviors indicative of anxiety, such as cowering. In future studies, however, it would be beneficial to quantify anxiety and determine, for example, whether individuals that are typically housed in social groups perform worse on cognitive challenges when tested alone than individuals that are typically housed alone.

Group Size as a Measure of Social Complexity. The use of social group size as a measure of social complexity has been criticized, because animals living in very large groups, such as ungulate herds, need not necessarily cope simultaneously with multiple types of differentiated relationships (61). However, in this study, we used published measures of species' average group size (27) as a proxy for social complexity, because Swanson et al. (27) found that this was no more or less effective than using the first principal component (PC1) from a principal component analysis of several different measures of social complexity in mammalian carnivores, indicating that group size is just as effective as more comprehensive measures of gregariousness available for mammalian carnivores. Furthermore, group size is a standard measure of social complexity that has been used repeatedly in the social complexity literature (34, 62, 63). For example, in a recent comparative study investigating the evolution of self-control (12), social group size was the predictor variable used to test the social complexity hypothesis. This study by MacLean et al. (12) is similar to our study in that they presented the same cognitive challenges to a number of individuals from a broad array of species and asked whether brain size, group size, or various ecological factors predicted success in these tests. We analyzed our data using group size as a continuous predictor variable (see below), but we also ran a separate analysis, in which we inquired whether social or nonsocial species (group size >1 or 1, respectively) were better able to open the puzzle box (results are shown in Table S7). These results confirmed the conclusion from our other analyses (Table 2) that group size or sociality does not influence performance in the puzzle box trials.

Total and Regional Brain Volumes. We obtained total brain volume (in milliliters) and adult body mass (in kilograms) data for each tested carnivore species from previously published datasets (46). Virtual endocasts, the digital casts made from the cranial bones comprising the brain case, were created using computed tomography (CT) and the software package MIMICS 11.02 (Materialise, Inc.) (table S7 in ref. 27). We obtained measures of total endocranial volume (in millimeters^3); the volumes of the anterior cerebrum, the posterior cerebrum, the total cerebrum (volumes of the posterior cerebrum and the anterior cerebrum), and the cerebellum plus brainstem, and body mass (in kilograms). Cranial and endocranial measures used in our analysis included combined cerebellum and brainstem volume in millimeters^3 (cerebellum plus brainstem), cerebrum anterior to the cruciate sulcus in millimeters^3 (volume of the anterior cerebrum), and cerebrum posterior to the cruciate sulcus in millimeters^3 (volume of the posterior cerebrum). The work by Swanson et al. (27) has additional details about these measurements, and the works by Sakai et al. (54, 55) have details on the CT scanning methods and application of MIMICS software to generate volumetric data for specific brain areas.

Statistical Methods. We defined a puzzle box trial as successful if the individual opened the box, and therefore, our response variable was binary (an individual did or did not open the puzzle box). Most individuals were tested in the puzzle box trials at least three times (mean = 4.2, median = 3, range = 1–10). Rather than averaging performance among these trials either within an individual or within a species, we used each individual trial as our unit of replication. We used two different datasets in our analysis of relationships between predictor traits and success opening the puzzle box. First, we used a large dataset containing puzzle box performance data from 495 trials on 140 different individuals in 39 species to investigate how total brain volume (from ref. 46), group size, dexterity, work time, and behavioral diversity affected performance in the puzzle box test. Second, we used a smaller dataset containing 209 trials on 65 individuals from 17 tested species to analyze the effects of size of specific brain regions on performance in the puzzle box trials (all data used in this study are in Dataset S1). Data documenting both total endocranial volume (overall brain size) and volumes of specific brain regions were obtained from CT scans from a total of 17 species (data from ref. 27).

To account for the shared evolutionary history among our subject species, we used Bayesian phylogenetic generalized linear mixed-effects models implemented in MCMCglmm (47, 48). Phylogeny was included as the inverse of the variance–covariance matrix, and we also included species as a random effect in these models (48). We obtained our mammalian phylogeny from an updated version of a recent mammalian supertree phylogeny (64) from the work by Fritz et al. (65). We pruned species not in Dataset S1 in R (version 3.2.0) (66) using the package geiger (version 2.0.3) (67). Pagel's λ is a measure of phylogenetic autocorrelation among the residuals that ranges from zero to one (51, 68). We estimated and report an approximate Bayesian equivalent of Pagel's λ (phylogenetic heritability) in all of our models as described elsewhere (48, 69).

We investigated how brain volume affected performance in the puzzle box task in a model that included the following fixed effects: brain volume, body mass, latency to approach the box, work time, behavioral diversity, manual dexterity, and average group size. In separate models, we included group size^2 but found no nonlinear relationship between group size and performance in the puzzle box task. We, therefore, did not include group size^2 in our other models or the results presented in the text or *SI Text*. We log-transformed (base e) brain volume, body mass, and all behavioral variables before analysis. We included individual identity as a random effect, because we had multiple measures on the same individuals. Although we tested multiple species at each zoo, we did not include a random effect for zoo identity, because we did not measure each species at each zoo. Inclusion of zoo identity as a random effect in preliminary analyses also failed to influence our inferences presented in the text or *SI Text* (results are presented above).

We used DIC (52) to examine the relative degree of fit of different models that we constructed to contain different combinations of the predictor variables (brain volume, body mass, behavioral variables, and group size) (Table 1). Overall, the model with the lowest DIC was the model containing the brain, body mass, behavioral, and group size predictor variables (Table 1), and therefore, we present the results from these analyses (Table 2).

We used DIC values from five candidate models to determine whether the volume of any specific brain region relative to total brain volume and body mass better predicted success in opening the puzzle box than total brain volume relative to body mass. These models contained (*i*) total brain volume relative to body mass, (*ii*) PC volume, (*iii*) AC volume, (*iv*) cerebrum volume, or (*v*) cerebellum plus brainstem volume. All of these models contained fixed effects for total brain volume and body mass (27) as well as random effects for species and individual identity. Because of our small sample sizes in these analyses and to avoid overparameterizing the models, we did not include the fixed effects for latency to approach the box, work time, behavioral diversity, manual dexterity, and average group size. The model with the lowest DIC value was assumed to be the top model.

For the random effects (species and individual identity) in all of our models, we used inverse Wishart priors that specified variance (V) and degree of belief in V (μ). Because our response variable was binary, we used a fixed prior ($V = 1$; fix = 1). For the random effect variance, we used weakly informative priors with a low degree of belief ($V = 1$; $\mu = 0.002$), which in this case, are also called inverse γ-priors and are widely used. Models for our first set of analyses on total brain volume (those shown in Table 2 and Tables S2–S8) were run for 8–10 million iterations with a burn-in of 150,000 and a thinning interval of 3,000. Models for our second set of analyses on brain region volume using 17 species were run for 85 million iterations with a burn-in of 250,000 and a thinning interval of 15,000 (Table S9). These different numbers of iterations were used to generate a minimum of an effective sample size of

>2,000 in all of the different models (but effective sample sizes were generally much higher than 2,000 as shown in Table 2 and Tables S2–S9).

We confirmed convergence of our posterior distributions using diagnostic tests available in the R package coda (version 0.17–1) (70), including the Geweke convergence diagnostic (71) and the Gelman–Rubin statistic (72). We ran all of our main models three times [although only results from the first Markov Chain Monte Carlo (MCMC) chain are shown] so that we could calculate the Gelman–Rubin statistic (potential scale reduction) (72–73). All potential scale reduction factors were $\leq$1.01, which is good evidence that the MCMC chains converged (72, 73). Autocorrelation among the MCMC chains can reduce the effective sample size, but in all of our models, our effective sample sizes were >2,000, indicating that our number of iterations and thinning interval were appropriate to generate sufficient sample sizes. All of our MCMCglmm models used slice sampling.

Inverse Wishart priors, like the ones that we used, are generally weakly informative, except when the posterior density is close to zero (74). In our models, the posterior density estimate for the random effect of species was relatively close to zero (diagnosed visually in trace plots). The use of more informative parameters or parameter-expanded priors may be used as an alternative in such situations (74). We, therefore, also examined how robust our results were to the use of alternative priors for the variance of the random effect for species. Specifically, we ran the same models using more informative priors (e.g., $V = 10$ and $\mu = 1$; $V = 100$ and $\mu = 1$, or $V = 100$ and $\mu = 2$) or parameter-expanded priors (e.g., $V = 1$; $\mu = 1$; $\alpha\mu = 0$; $\alpha V = 625$) (75). Using these different priors for the variance of the random effect for species, the models once again provided the inference that species with larger brains for their body mass were better able to open the puzzle box, although the posterior distributions for the parameters changed. Moreover, if we reran the models excluding the random effect of species entirely, we still found the same qualitative results that species with larger brains for their body mass were better able to open the puzzle box. This result suggests that our use of a weak prior ($V = 1$; $\mu = 0.002$) for the variance for the random effect of species was appropriate and that the results were robust to the use of alternative priors.

Table S1. Summary of the number of mammalian carnivore species and different individuals within each species in which we examined their performance in opening up the puzzle box

Family	No. of species	Relative brain size	Total no. of trials	No. of individuals tested	No. of individuals successful	Individuals successful (%)
Ailuridae	1	0.189	18	4	1	25
Canidae	7	0.193 (−0.13–0.41)	75	25	6	24
Felidae	13	−0.093 (−0.36–0.23)	179	52	15	28.8
Herpestidae	2	−0.417 (−0.59 to −0.24)	15	5	0	0
Hyaenidae	2	−0.192 (−0.26 to −0.12)	23	7	2	28.6
Mustelidae	4	−0.005 (−0.37–0.14)	81	17	8	47.1
Procyonidae	3	0.077 (0.02–0.18)	42	13	7	53.8
Ursidae	5	0.196 (−0.16–0.94)	45	13	9	69.2
Viverridae	2	−0.153 (−0.24 to −0.06)	17	4	1	25
Sum or mean	39	—	495	140	49	35

Relative brain size (mean and range) for each family is the residual brain volume from a general linear model containing log-transformed brain volume (kilograms) as the response variable and log-transformed brain volume (endocranial volume in milliliters) as the predictor variable and is for presentation purposes only.

Table S2. Results from Bayesian phylogenetic generalized linear mixed effects models to investigate the predictors of success in opening the puzzle box in 39 mammalian carnivore species

Fixed or random effect	Effective sample size	Posterior mean (95% CI)	Posterior mode	pMCMC
Random effect				
Species	2,263	20.9 (0.001–56.5)	8.2	—
Individual identification	2,617	19.3 (6.3–34.5)	14.1	—
Fixed effect				
Intercept	2,617	−18.3 (−30.4 to −7.8)	−18.2	0.0004
Brain volume	2,617	1.4 (−0.6–3.4)	1.2	0.14
Latency to approach	2,617	−0.14 (−0.5–0.2)	−0.1	0.52
Work time	2,713	0.3 (−0.06–0.7)	0.3	0.09
Behavioral diversity	2,617	1.3 (−2.5–5.7)	1.2	0.52
Dexterity	2,617	2.7 (−0.4–6.1)	2.5	0.07
Group size	2,617	−0.09 (−0.4–0.2)	−0.08	0.54

This model investigates the effect of absolute brain volume without body mass in the model (Table 2 shows the same model including body mass). pMCMC is the Bayesian *P* value. Sample sizes are 495 trials on 140 individuals from 39 different species.

Table S3. Results from Bayesian phylogenetic generalized linear mixed-effects models to investigate the predictors of success in opening the puzzle box in 23 mammalian carnivore species when at least three individuals were tested in each species

Fixed or random effect	Effective sample size	Posterior mean (95% CI)	Posterior mode	pMCMC
Random effect				
Species	3,604	27.3 (0.0006–79.3)	8.1	—
Individual identification	3,811	22.8 (6.3–46.6)	15	—
Fixed effect				
Intercept	3,906	−57.2 (−97.3 to −21.3)	−54.6	0.0004
Brain volume*	4,211*	16.2 (3.6–30.5)*	15.1*	0.007*
Body mass*	4,567*	−9.5 (−18.6 to −1.6)*	−8.9*	0.011*
Latency to approach	4,950	−0.1 (−0.5–0.4)	−0.1	0.7
Work time*	4,950*	0.6 (0.2–1.1)*	0.6*	0.009*
Behavioral diversity	4,950	1.1 (−3.7–6.3)	0.7	0.64
Dexterity	4,671	3 (−0.5–7.4)	2.8	0.11
Group size	4,950	0.01 (−0.4–0.4)	0.03	0.92

pMCMC is the Bayesian *P* value. Sample sizes are 398 trials on 112 individuals from 23 different species.
*Statistically significant.

Table S4. Results from Bayesian phylogenetic generalized linear mixed effects models to investigate the predictors of success in opening the puzzle box in 18 mammalian carnivore species when at least four individuals were tested in all species

Fixed or random effect	Effective sample size	Posterior mean (95% CI)	Posterior mode	pMCMC
Random effect				
Species	2,634	34.3 (0.0006–102.6)	7.6	—
Individual identification	2,647	21.2 (4.9–44.7)	13.2	—
Fixed effect				
Intercept	2,456	−59.7 (−106 to −16.3)	−53.6	0.0012
Brain volume*	2,649*	16.6 (1.8–32.6)*	14.6*	0.012*
Body mass*	2,707*	−9.7 (−20.1 to −1.3)*	−9.3*	0.018*
Latency to approach	3,284	0.07 (−0.4–0.6)	0.1	0.81
Work time*	3,284*	0.63 (0.1–1.1)*	0.6*	0.01*
Behavioral diversity	3,284	1.47 (−3.6–6.5)	0.7	0.55
Dexterity	3,284	2.85 (−0.9–6.9)	3.1	0.13
Group size	3,284	0.08 (−0.4–0.5)	0.1	0.76

pMCMC is the Bayesian *P* value. Sample sizes are 348 trials on 97 individuals from 18 different species.
*Statistically significant.

Table S5. Results from Bayesian phylogenetic generalized linear mixed effects models to investigate the predictors of success in opening the puzzle box in 39 mammalian carnivore species when at least 3 trials were performed in all individuals (total number of trials per individual ranged from 3 to 10)

Fixed or random effect	Effective sample size	Posterior mean (95% CI)	Posterior mode	pMCMC
Random effect				
Species	4,518	15 (0.001–41.8)	4.9	—
Individual identification	3,732	20.3 (7-38.8)	16.04	—
Fixed effect				
Intercept	4,153	−35.1 (−58.1 to −12.7)	−30.9	0.002
Brain volume*	4,702*	8.05 (0.79–16.02)*	8.3*	0.03*
Body mass	4,803	−4.3 (−9.23–0.01)	−3.73	0.06
Latency to approach	5,431	−0.11 (−0.54–0.3)	−0.17	0.61
Work time	5,200	0.33 (−0.04–0.74)	0.29	0.08
Behavioral diversity	5,200	2.06 (−1.7–6.3)	1.53	0.3
Dexterity	4,716	2.4 (−0.61–5.5)	1.85	0.11
Group size	4,912	−0.08 (−0.39–0.21)	−0.02	0.59

pMCMC is the Bayesian *P* value. Sample sizes are 483 trials on 131 individuals from 39 different species.
*Statistically significant.

Table S6. Results from Bayesian phylogenetic generalized linear mixed effects models to investigate the predictors of success in opening the puzzle box in 39 mammalian carnivore species when we restricted our analysis to examining only the first three trials for each individual that was tested at least three times

Fixed or random effect	Effective sample size	Posterior mean (95% CI)	Posterior mode	pMCMC
Random effect				
Species	2,633	19.6 (0.0002–56.3)	0.1	—
Individual identification	2,648	18.6 (4.4–37.9)	13.2	—
Fixed effect				
Intercept	3,012	−35.6 (−62.2 to −11.9)	−30.9	0.002
Brain volume	3,585	7.3 (−0.03–16.1)	7.7	0.052
Body mass	3,522	−3.5 (−8.1–1.8)	−3.6	0.13
Latency to approach	3,284	−0.19 (−0.7–0.3)	−0.2	0.49
Work time	3,284	0.25 (−0.3–0.8)	0.2	0.35
Behavioral diversity	3,284	2.8 (−1.5–7.1)	3.1	0.18
Dexterity	3,284	2.7 (−0.7–5.9)	2.1	0.08
Group size	3,123	−0.06 (−0.4–0.2)	−0.04	0.7

pMCMC is the Bayesian *P* value. Sample sizes are 388 trials on 131 individuals from 39 different species.

Table S7. Results from Bayesian phylogenetic generalized linear mixed effects models to investigate the predictors of success in opening the puzzle box in 39 mammalian carnivore species

Fixed or random effect	Effective sample size	Posterior mean (95% CI)	Posterior mode	pMCMC
Random effect				
Species	4,529	16.4 (0.0002–40)	4.9	—
Individual identification	3,398	21.1 (7.6–40.5)	16	—
Fixed effect				
Intercept	5,200	−1,158 (−110,700–110,400)	−3,215	0.99
Brain volume*	4,876*	8.1 (0.47–10.6)*	7.02*	0.03*
Body mass	5,200	−4.4 (−8.9–0.62)	−3.6	0.06
Latency to approach	5,200	−0.13 (−0.57–0.28)	−0.15	0.56
Work time	4,583	0.33 (−0.05–0.72)	0.35	0.09
Behavioral diversity	5,200	1.6 (−2.5–5.8)	1.94	0.41
Dexterity	4,302	2.8 (−0.35–6)	2.3	0.064
Social species	5,200	1,122 (−110,400–110,700)	3,168	0.99

In this model, group size is a binary variable, where species are either solitary (group size of 1) or social (group size >1). Reference value in the intercept is social species. pMCMC is the Bayesian *P* value. Sample sizes are 495 trials on 140 individuals from 39 different species.

*Statistically significant.

Table S8. In nine of the puzzle box trials, individuals opened the door to retrieve the reward but did not eat the food reward

Fixed or random effect	Effective sample size	Posterior mean (95% CI)	Posterior mode	pMCMC
Random effect				
Species	2,426	14.4 (0.0007–42.9)	3.8	—
Individual identification	2,235	21.1 (6.4–39.1)	16.1	—
Fixed effect				
Intercept	2,652	−36.4 (−60.8 to −15.6)	−38.5	0.0008
Brain volume*	2,734*	8.5 (1.5–16.2)*	7.6*	0.011*
Body mass*	2,795*	−4.6 (−9.4 to −0.4)*	−4.1*	0.035*
Latency to approach	2,617	−0.13 (−0.53–0.3)	−0.06	0.55
Work time	2,597	0.33 (−0.05–0.71)	0.3	0.08
Behavioral diversity	2,617	1.7 (−2.2–5.8)	1.9	0.38
Dexterity	2,617	2.7 (−0.4–5.9)	1.6	0.07
Group size	2,617	−0.04 (−0.34–0.25)	1.3	0.8

The behavior of these nine animals may suggest differences among individuals in their motivation to open the puzzle box. We performed the same statistical model shown in Table 2 but excluded these nine trials. Results from Bayesian phylogenetic generalized linear mixed effects models to investigate the predictors of success in opening the puzzle box in 39 mammalian carnivore species are shown. pMCMC is the Bayesian *P* value. Sample sizes are 486 trials on 140 individuals from 39 different species.

*Statistically significant.

Table S9. Results from Bayesian phylogenetic generalized linear mixed effects models to investigate whether specific brain region volumes predicted success in opening the puzzle box in 17 mammalian carnivore species

Brain region and fixed/random effects	Effective sample size	Posterior mean	Posterior mode	95% CI	pMCMC
Anterior cerebrum					
Species	5,381	18.7	−0.21	0.0002–80.3	—
Individual identification	2,714	107.8	44.9	8.6–308.6	—
Fixed effects					
Intercept	4,652	−136.3	−93.3	−411.7–73.1	0.17
Brain volume	4,939	10.9	5.4	−12.6–39.5	0.32
Body mass	4,500	−8.7	−6.3	−25.1–3.2	0.13
Anterior cerebrum volume	5,130	3.1	1.8	−3.9–11.4	0.31
Posterior cerebrum					
Species	5,650	12.1	−0.12	0.0002–47.7	—
Individual identification	3,152	101.3	42.1	9.4–280.3	—
Intercept	5,015	−147.3	−83.9	−429.5–56.2	0.12
Brain volume*	4,587*	38.1*	31.2*	−5.6–89.4*	0.04*
Body mass	5,146	−8.1	−4.2	−24.3–3.5	0.14
Posterior cerebrum volume	4,780	−24.7	−14.5	−68.3–10.6	0.12
Total cerebrum					
Species	5,571	18.3	−0.19	0.0002–78.5	—
Individual identification	3,084	105	46	12.9–294	—
Intercept	4,252	−155.4	−143.3	−436.6–73.3	0.13
Brain volume	5,650	22.3	16.3	−137.2–184.8	0.76
Body mass	4,222	−9.2	−6.8	−24.6–3.9	0.12
Total cerebrum volume	5,650	−7.1	5.9	−164.2–145.2	0.92
Brainstem and cerebellum					
Species	4,616	27.1	−0.24	0.0002–110.2	—
Individual identification	2,760	109	52.8	11.6–304.7	—
Intercept	3,564	−158.8	−82.5	−456.2–64.1	0.11
Brain volume	5,008	17.5	10.9	−23.5–61.9	0.33
Body mass	3,863	−9.1	−6.9	−26.4–4.4	0.13
Brainstem and cerebellum volume	5,986	−2.3	0.3	−40.7–31.9	0.92
Brain volume					
Species	4,847	9	−0.10	0.0002–41.5	—
Individual identification	4,149	80.2	32.3	8.7–205.2	—
Intercept	4,644	−137.3	−79.3	−377.6–31.04	0.09
Body mass	4,669	−8.2	−6.6	−21.2–2.2	0.09
Brain volume	4,911	13.6	9.1	−4.1–37.5	0.11

pMCMC is the Bayesian *P* value.

*Statistically significant.

Movie S1. Examples of the behavior of 19 different carnivore species when interacting with the puzzle box. The species represented are listed in order of appearance: Red Panda (*Ailurus fulgens*), Binturong (*Arctictis binturong*), Grizzly Bear (*Ursus arctos*), White-nosed Coati (*Nasua narica*), Bobcat (*Lynx rufus*), Arctic Fox (*Alopex lagopus*), Polar Bear (*Ursus maritimus*), Black Bear (*Ursus americanus*), Snow Leopard (*Panthera uncia*), Pallas cat (*Otocolobus manul*), Striped Hyena (*Hyaena hyaena*), Spotted Hyena (*Crocuta crocuta*), American Badger (*Taxidea taxus*), Leopard (*Panthera pardus*), Amur Tiger (*Panthera tigris*), River Otter (*Lontra canadensis*), Maned Wolf (*Chrysocyon brachyurus*), Wild Dog (*Lycaon pictus*), and Wolverine (*Gulo gulo*).

Movie S1

Other Supporting Information Files

Dataset S1 (XLSX)