tbd...

Supplementary material

tbd...

Contents

Overview

... Next we describe the different tasks we sused ...

tasks - stability, reliability

predictors - predictability

Methods

Participants

A total of 41 great apes participated at least once in one of the tasks. This included 8 Bonobos (3 females, age 7.3 to 38.2), 22 Chimpanzees (17 females, age 2.6 to 55.2), 6 Gorillas (4 females, age 2.7 to 21.9), and 6 Orangutans (4 females, age 17 to 40.5). The sample size at the different time points ranged from 0 to 21. Figure S1 visualizes the sample size across time points. We tried to test all apes at all time points but this was not always possible due to a lack of motivation or construction works. All apes participate in cognitive

research on a regular basis. Many of them have ample experience with the very tasks we used in the current study.

Apes were housed at the Wolfgang Köhler Primate Research Center located in Zoo Leipzig, Germany. They lived in groups, with one group per species and two chimpanzee groups. Research was noninvasive and strictly adhered to the legal requirements in Germany. Animal husbandry and research complied with the European Association of Zoos and Aquaria Minimum Standards for the Accommodation and Care of Animals in Zoos and Aquaria as well as the World Association of Zoos and Aquariums Ethical Guidelines for the Conduct of Research on Animals by Zoos and Aquariums. Participation was voluntary, all food was given in addition to the daily diet, and water was available ad libitum throughout the study. The study was approved by an internal ethics committee at the Max Planck Institute for Evolutionary Anthropology.

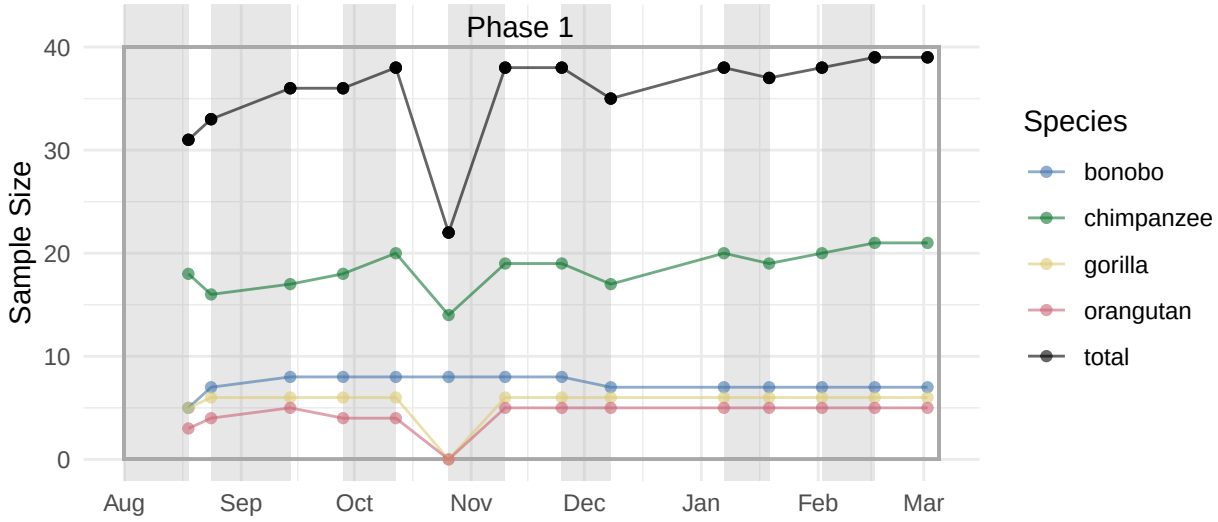


Figure S1: Sample size by species across the different time points. Time point specific predictor variables were collected during the time between two time points (shaded regions) to predict the next.

Setup

Apes were tested in familiar sleeping or observation rooms by a single experimenter. Whenever possible, they were tested individually. The basic setup comprised a sliding table positioned in front of a clear Plexiglas panel with three holes in it. The experimenter sat on a small stool and used an occluder to cover the sliding table (see Figure S2).

Tasks

The tasks we selected are based on published procedures and are commonly used in the field of comparative psychology. The original publications often include control conditions to rule out alternative, non-cognitive explanations. We did not include such controls here and only ran the experimental conditions. For each task, we refer to the publication we used to model our procedure. We ask the reader to read these papers if they want to know more about control conditions and/or a detailed discussion of the nature of the underlying cognitive mechanisms.

Example videos for each task can be found in the associated online repository in `videos/`.

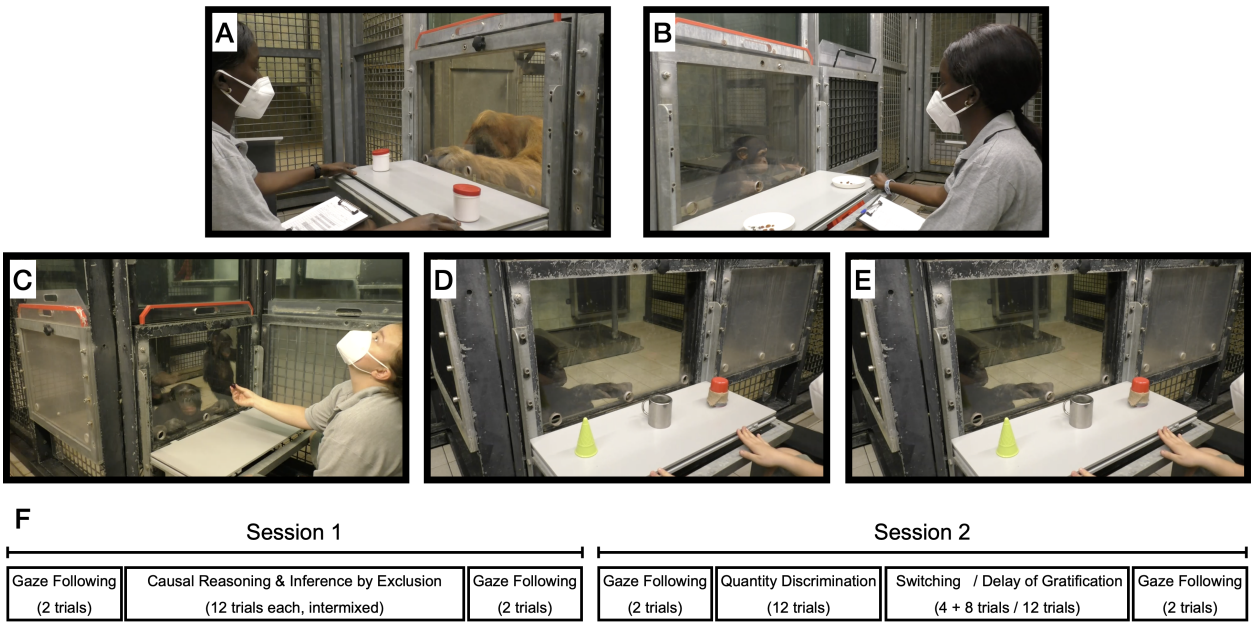


Figure S2: Setup used for the six tasks. A) Causal reasoning and inference by exclusion. B) Quantity discrimination. C) Gaze following. D) Switching. E) Delay of gratification.

Causal inference

The causal inference task was modeled after Call (2004). Two identical cups with a lid were placed left and right on the table (Figure S2A). The experimenter covered the table with the occluder, retrieved a piece of food, showed it to the ape, and hid it in one the cups outside the participant's view. Next, the experimenter removed the occluder, picked up the baited cup and shook it three times, which produced a rattling sound. Next, the cup was put back in place, the sliding table pushed forwards, and the participant made a choice by pointing to one of the cups. If they picked the baited cup, their choice was coded as correct, and they received the reward. If they chose the empty cup, they did not. Participants received 12 trials. The location if the food was counterbalanced; 6 times in the right cup and 6 times in the left. Causal inference trials were intermixed with inference by exclusion trials.

We assume that apes locate the food by reasoning that the food – a solid object – caused the rattling sound and must thus be in the shaken cup.

Inference by exclusion

Inference by exclusion trials were also modeled after Call (2004) and followed a very similar procedure compared to causal inference trials. After covering the two cups with the occluder, the experimenter placed the food in one of the cups and covered both with the lid. Next, they removed the occluder, picked up the empty cup and shook it three times. In contrast to the causal inference trials, this did not produce any sound. The experimenter then pushed the sliding table forward and the participant made a choice by pointing to one of the cups. Correct choice was coded when the baited (non-shaken) cup was chosen. If correct, the food was given to the ape. There were 12 inference by exclusion trials, intermixed with causal inference trials. The order was counterbalanced: 6 times the left cup was baited, 6 times the right.

We assume that apes reason that the absence of a sound suggests that the shaken cup is empty. Because they saw a piece of food being hidden, they exclude the empty cup and infer that the food is more likely to be in the non-shaken cup.

Gaze Following

The gaze following task was modeled after Brauer, Call, and Tomasello (2005). The experimenter sat opposite the ape and handed over food at a constant pace. That is, the experimenter picked up a piece of food, briefly held it out in front of her face and then handed it over to the participant. After a predetermined (but varying) number of food items had been handed over, the experimenter again picked up a food item, held it in front of her face and then looked up (i.e., moving her head up - see Figure S2C). The experimenter looked to the ceiling, no object of particular interest was placed there. After 10s, the experimenter looked down again, always handed over the food and the trial ended. We coded whether the participant looked up during the 10s interval.

We assume that participants look up in order to follow the experimenter's gaze to locate a potentially noteworthy object.

Quantity discrimination

For this task, we followed the general procedure of Hanus and Call (2007). Two small plates were presented left and right on the table (see Figure S2B). The experimenter covered the plates with the occluder and placed 5 small food pieces on one plate and 7 on the other. Then they pushed the sliding table forwards, and the participant made a choice. We coded as correct when the subject chose the plate with the larger quantity. Participants always received the food from the plate they chose. There were 12 trials, 6 with the larger quantity on the right and 6 on the left (order counterbalanced).

We assume that ???

Switching

This task was modeled after Haun et al. (2006). Three differently looking cups (metal cup with handle, red plastic ice cone, red cup without handle - Figure S2D) were placed next to each other on the table. There were two conditions. In the place condition, the experimenter hid a piece of food under one of the cups in full view of the participant. Next, the cups were covered by the occluder and the experimenter switched the position of two cups, while the reward remained in the same location. Next, the experimenter removed the occluder and pushed the table forward. We coded as correct if the participant chose the location where the food was hidden. Participants received four trials in this condition.

The place condition was run first. The feature condition followed the same procedure, but now the experimenter also moved the reward when switching the cups. The switch between conditions happened without informing the participant in any way. A correct choice in this condition meant choosing the location to which the cup plus the food were moved. Here, participants received eight trials.

The dependent measure of interest for this task was calculated as: `[proportion correct place] - (1 - [proportion correct feature])`. Positive values in this score mean that participants could quickly switch from choosing based on location to choosing based on feature. High negative values suggest that participants did not or hardly switch strategies.

Based on the results of Haun et al. (2006), we assume that apes have a tendency to expect the food to remain in the same location. When this strategy is no longer successful in the feature trials, they have to switch strategies and try a different one.

Delay of gratification

tbd.

Data collection

One time point meant running all tasks with all participants. Within each time point, the tasks were organized in two sessions (see Figure S2F). Session 1 started with 2 gaze following trials. Next was a pseudo random mix of causal inference and inference by exclusion with 12 trials per task but no more than two trials of the same task in a row. At the end of session 1, there were again 2 gaze following trials. Sessions 2 also started with 2 gaze following trials, followed by quantity discrimination and switching. Finally, there were again 2 gaze following trials. B spreading out or mixing tasks we hoped to keep subjects more attentive and engaged.

The order of tasks was the same for all subjects. So was the positioning of food items within each task. The counterbalancing can be found in the coding sheets in the online repository in `documentation/` [**to be added**]. This exact procedure was repeated at each time point so that the results would be comparable across participants. The two sessions were usually spread out across two adjacent days. For the larger chimpanzee group, they were sometimes spread out across 4 days.

The interval between two time points was planned to be two weeks. However, it was not always possible to follow this schedule so that that some intervals are longer/shorter. Figure S1 visualizes the intervals between time points.

We collected data in two phases. Phase 1 started on August 1st, 2020, lasted until March 5th, 2021 and included 14 time points (see Figure S1). Phase 2 started on , lasted until and had time points.

Predictors

In addition to the data from the cognitive tasks, we collected data for a range of predictor variables. The goal here was to find variables that are systematically related to inter- and/or intra-individual variation in cognitive performance. That is, we were interested to see which variables allow us to predict cognitive performance. The second part of the analysis section, describes the method we used to determine the predictive value of each variable.

Predictors could either vary with the individual (stable individual characteristics; e.g. sex or rearing history), vary with individual and time point (variable individual characteristics; e.g. sickness or sociality), vary with group membership (group life; e.g. time spent outdoors or disturbances) or vary with the testing arrangements (testing arrangements; e.g. presence of an observer or participation in other tests).

Most predictors were collected via a diary that the animal caretakers filled out on a daily basis. Here, the caretakers were asked a range of questions about the presence of a predictor and its severity. The diary (in German) can be found in `documentation/` in the associated online repository.

Stable individual characterisitcs

These predictors are stable individual differences. As a source, we used the ape handbook at Zoo Leipzig. Figure S3 gives an overview of the distribution of the different characteristics in the sample.

Age Absolute age of the individual. For some older individuals, only the year of birth was known. In these cases we calculated age with January 1st of that year as the birthday.

Sex Participant's biological sex.

Rearing history Here, we differentiated between, `mother-reared`, `hand-reared` and `unknown`. The last category was used only for three chimpanzees. In the analysis, we classified them as `hand-reared` to facilitate model fitting (i.e. it is very difficult to estimate a parameter for a factor level with so little data). We think

this decision is justified because the individuals in question have spent most of their life in close contact to humans and not in a larger chimpanzee group.

Time lived in Leipzig Absolute time the individual has lived in Leipzig Zoo. All apes living in Leipzig are involved in behavioral research. Thus, we take this measure to be a rough proxy of how much experience an individual has had with cognitive research.

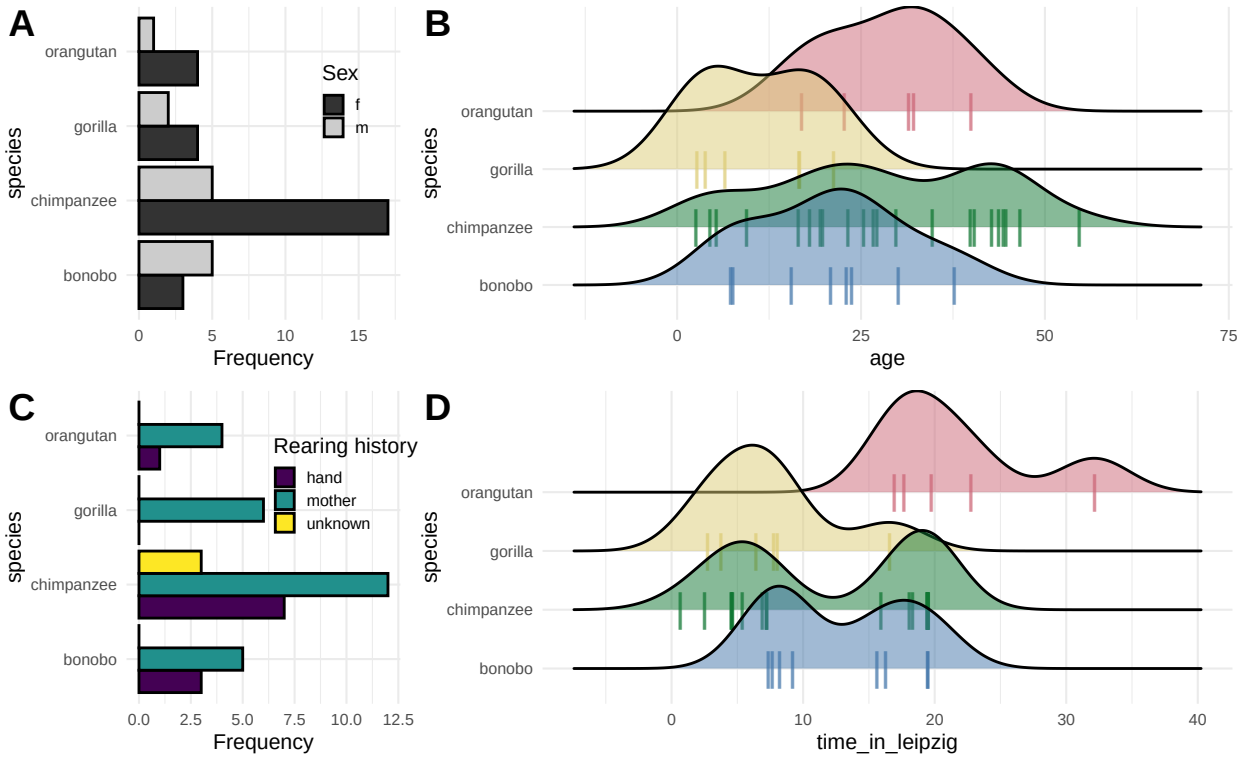


Figure S3: Stable individual characteristics. A) participant sex, B) age distribution by species, C) rearing history, D) time lived in leipzig by species.

Variable individual characteristics

These predictors varied by participant and time point.

Rank We asked caretakers to order individuals within a given group for their rank. Ties were allowed. This was done at each time point. An individual's rank was mostly stable (see Figure **??**A) across time points, however, there was some variation.

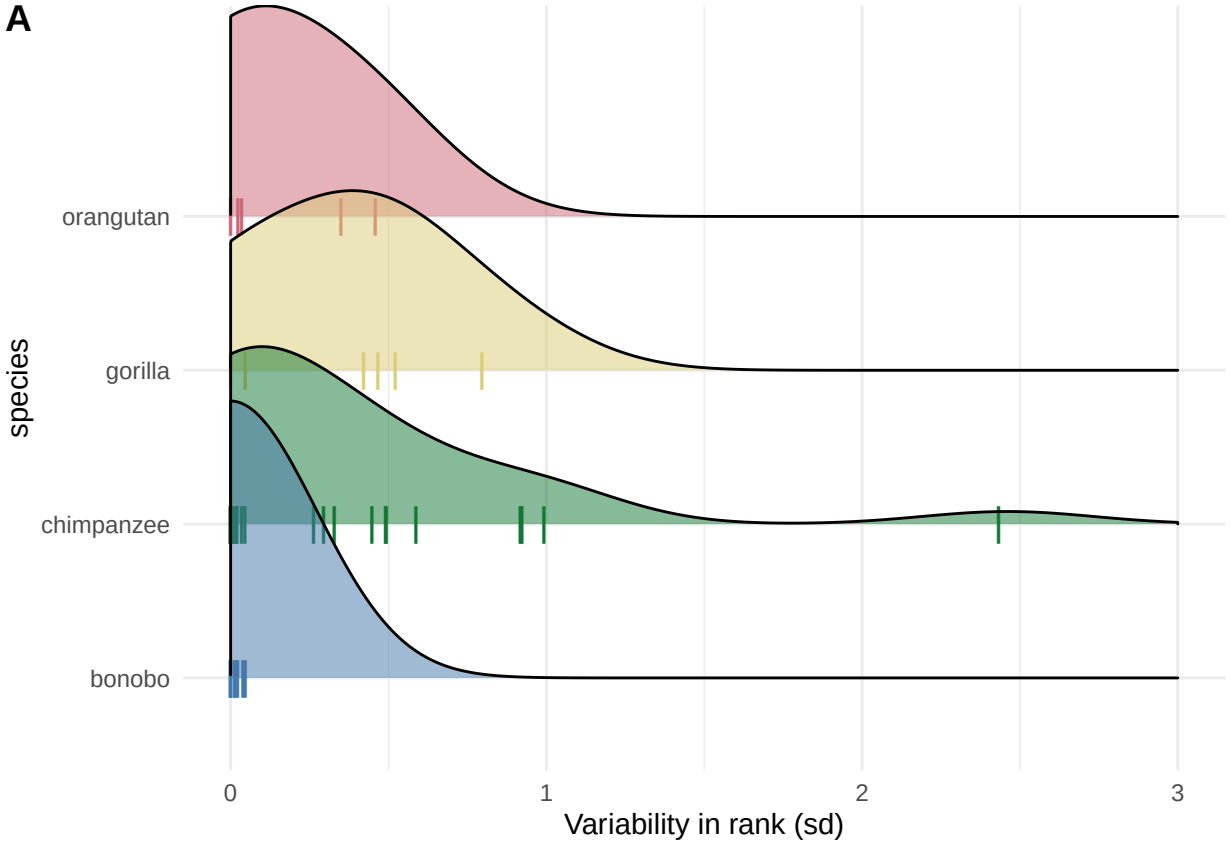
Sickness As part of the caretakers' daily diary, we asked whether an individual was sick and if yes, how sever the sickness was on a scale from 1 to 7. For each time point, we used the mean of the daily sickness ratings as predictor.

Sociality We conducted proximity scans for all groups in the early afternoon on every work day (Monday to Friday). That is, we expect 10 scans for each time point. For each individual, we recorded which individuals are within arms reach. Research assistants used a tablet to record their observations.

To derive individual specific estimates of sociality for each time point, we fit a variant of a Social Relations Model (Snijders and Kenny 1999) to the proximity data. These models allow estimating an individual specific sociality index while accounting for the dyadic nature of social interaction. Social relations model usually deal with directed behaviors (e.g. individual i is grooming individual j). Because the behavior we observed was symmetric, we cannot differentiate between the actor and receiver. Kajokaite et al. (2021) suggested to speak of a Multiple Membership Relations Model (see also Leckie 2019) in such a context, which simply estimates how likely likely an individual is to be observed in proximity to another individual.

In `brms` syntax, our model had the following structure: `count | trials(n) ~ group + (time_point | mm(focal, associates)) + (time_point | dyad)`. The dependent variable `count | trials(n)` is the number of times a dyad has been observed (`count`) at a time point relative to the number of scans taken for that time point (`trials(n)`). The fixed effect `group` estimates group difference in sociality. The random effect `(time_point | mm(focal, associates))` estimates the sociality for each individual. In that, the multi-membership grouping term `mm(focal, associates)` captures the fact that the assignment of the two roles (focal and associate) is arbitrary in the context of a symmetric behavior. The random slope `time_point` (treated as a factor) allowed us to estimate sociality for each time point. Finally, the random effect `(time_point | dyad)` accounts for dyad composition; in some cases a particular dyad composition (e.g. mother and infant) might be sufficient to explain high levels of sociality in an individual.

For each individual and time point, we extracted the sociality estimates and used them to predict cognitive performance in the different tasks for that time point. Figure **??**B visualizes the sociality measures for one group across the different time points.



Group life

This set of predictors varied by time point and group, but were the same for all individuals in that group. They were recorded in the animal caretaker diary. Figure S4 visualizes the different variables across time

points.

Time outdoors Each day, the animal caretakers noted in the diary how many hours each group spent in the outdoor enclosure. To compute the predictor, we averaged across these values for each time point and group.

Disturbances The animal caretakers also noted down if there were any unusual disturbances for a particular group. Example were construction works in the building, heavy weather conditions or green-keeping activities. In addition, the caretakers rated how disturbing they judged these events to be on a scale from 1 to 7. For each time point, we calculated the mean of these ratings.

Life events This variable captured whether there were any notable events within the group. Examples were fights in the group or the temporal removal of some individuals for medical procedures. Again, we asked the caretakesr to rate the severity of these events and averaged across them.

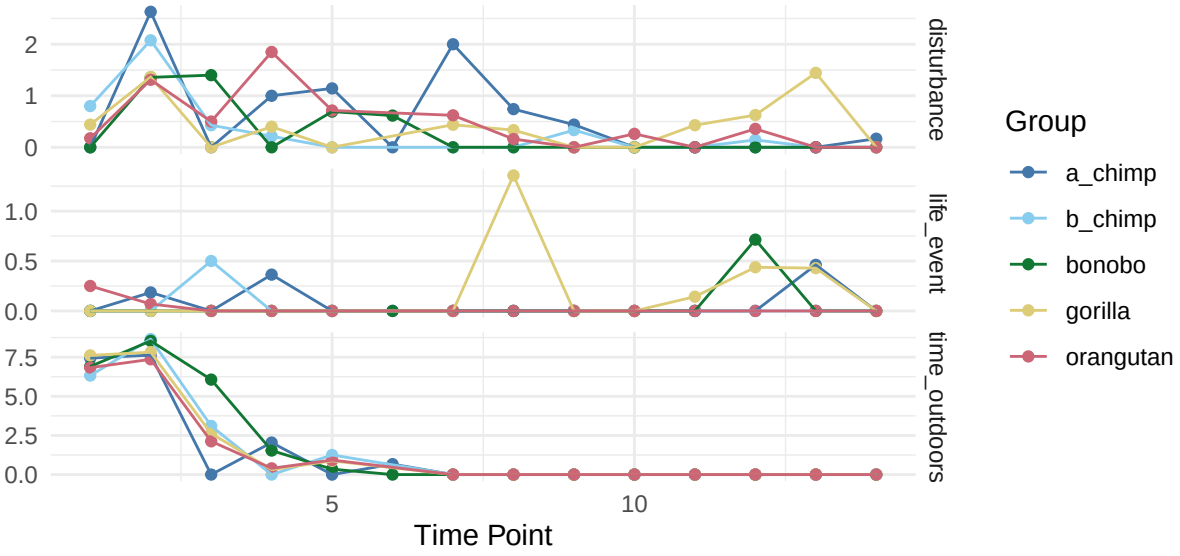


Figure S4: Variation in group life related measures across groups and time points.

Testing arrangements

Testing arrangements varied between individuals, sessions and time points. The experimenter recorded them either based on their observations during testing or from the testing schedule which lists all studies along with their participants that take place on a particular day.

Observer We noted whether or not there was another animal in the same room or the room adjacent to the one the participant was in.

Study on same day This predictor recorder whether or not the participant had already participated in a different test on the same day. The experimenter took this information from the testing schedule.

Studies since last time point Here we counted in how many other studies the participant had taken part since the last time they were tested in that particular task. The experimenter took this information from the testing schedule.

Analytical framework

Structural equation modelling

Stability, reliability, relations between tasks

Simulations

Projection predictive inference

predictor selection

Results

Phase 1

Stability

Reliability

Relations between tasks

Predictability

Summary

Appendix

Brauer, Juliane, Josep Call, and Michael Tomasello. 2005. "All Great Ape Species Follow Gaze to Distant Locations and Around Barriers." *Journal of Comparative Psychology* 119 (2): 145.

Call, Josep. 2004. "Inferences About the Location of Food in the Great Apes (Pan Paniscus, Pan Troglodytes, Gorilla Gorilla, and Pongo Pygmaeus)." *Journal of Comparative Psychology* 118 (2): 232.

Hanus, Daniel, and Josep Call. 2007. "Discrete Quantity Judgments in the Great Apes (Pan Paniscus, Pan Troglodytes, Gorilla Gorilla, Pongo Pygmaeus): The Effect of Presenting Whole Sets Versus Item-by-Item." *Journal of Comparative Psychology* 121 (3): 241.

Haun, Daniel BM, Josep Call, Gabriele Janzen, and Stephen C Levinson. 2006. "Evolutionary Psychology of Spatial Representations in the Hominidae." *Current Biology* 16 (17): 1736–40.

Kajokaite, Kotrina, Andrew Whalen, Jeremy Koster, and Susan Perry. 2021. "Fitness Benefits of Providing Services to Others: Grooming Predicts Survival in a Neotropical Primate." *bioRxiv*. https://doi.org/10.1101/2020.08.04.235788.

Leckie, George. 2019. "Multiple Membership Multilevel Models." http://arxiv.org/abs/1907.04148.

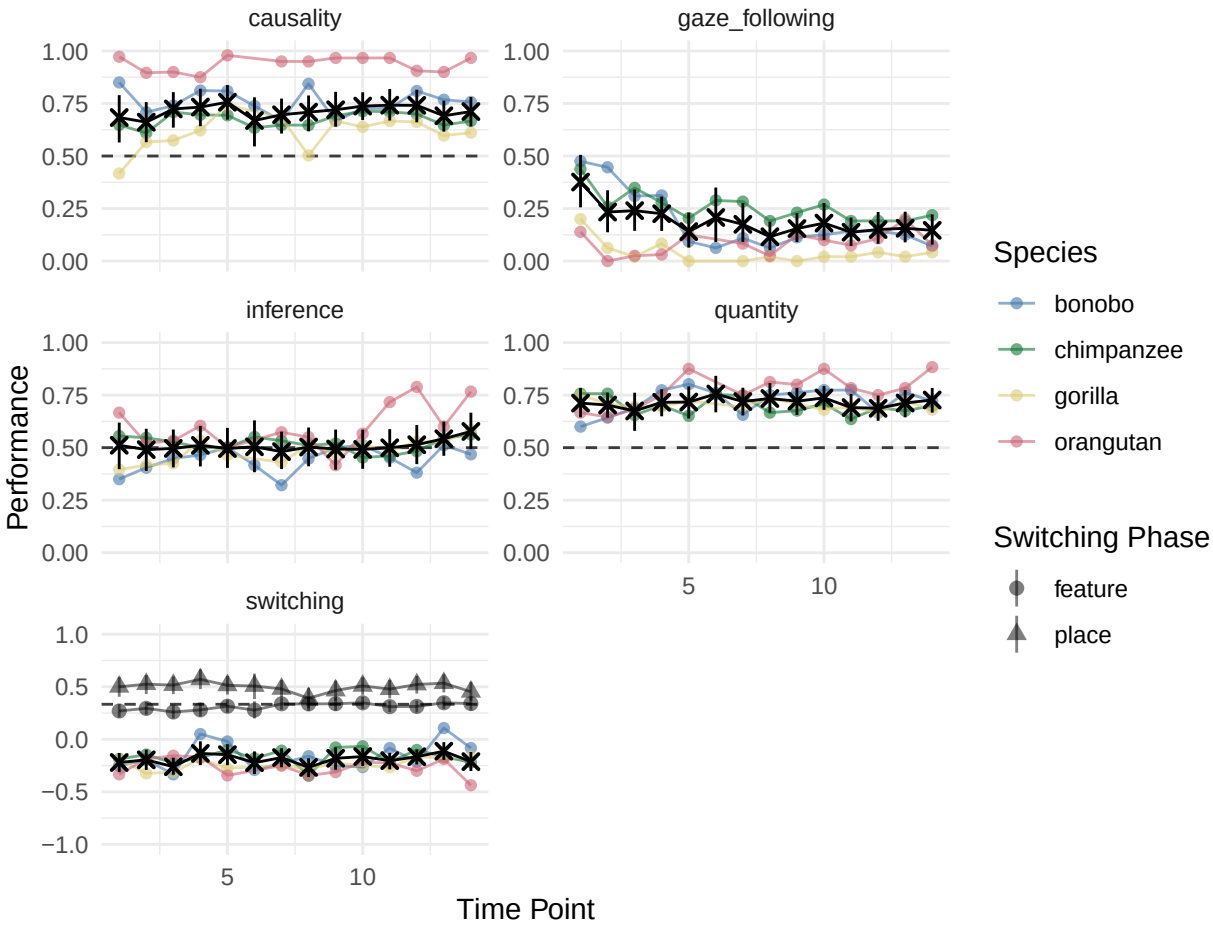


Figure S5: Results from the five cognitive tasks across time points. Black crosses show mean performance at each time point across species (with 95% CI). Colored dots show mean performance by species. Dashed line shows the chance level whenever applicable. The panel for switching includes triangles and dots showing the mean performance in the two phases from which the overall performance score was computed.

Snijders, Tom AB, and David A Kenny. 1999. "The Social Relations Model for Family Data: A Multilevel Approach." *Personal Relationships* 6 (4): 471–86.