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.

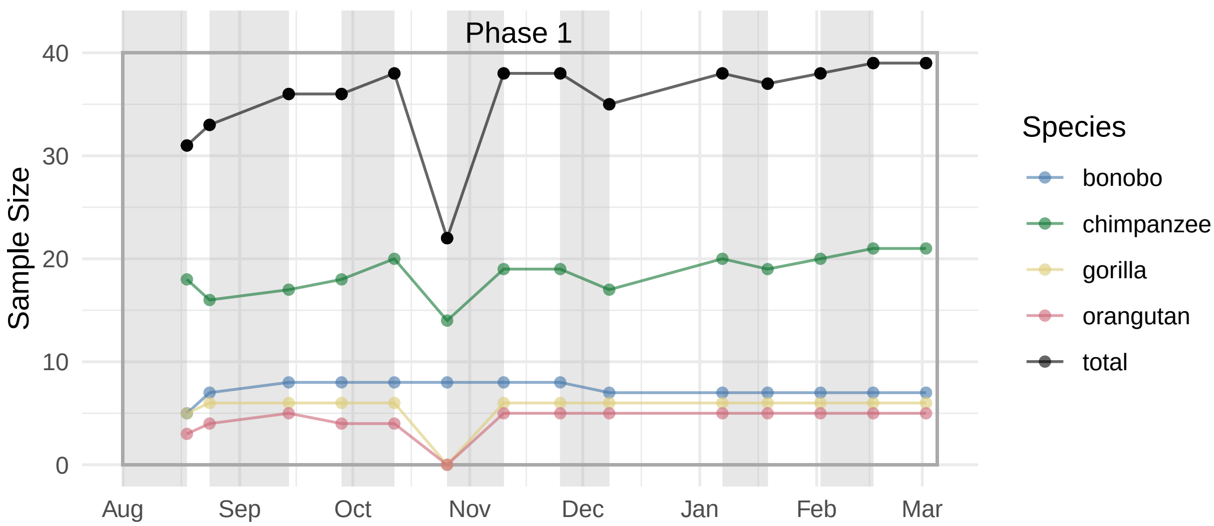

Figure S1: Sample size by species across the different time points. Time point specific predictor variables were collected during the time between two time points (shaded regions) to predict the next.

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

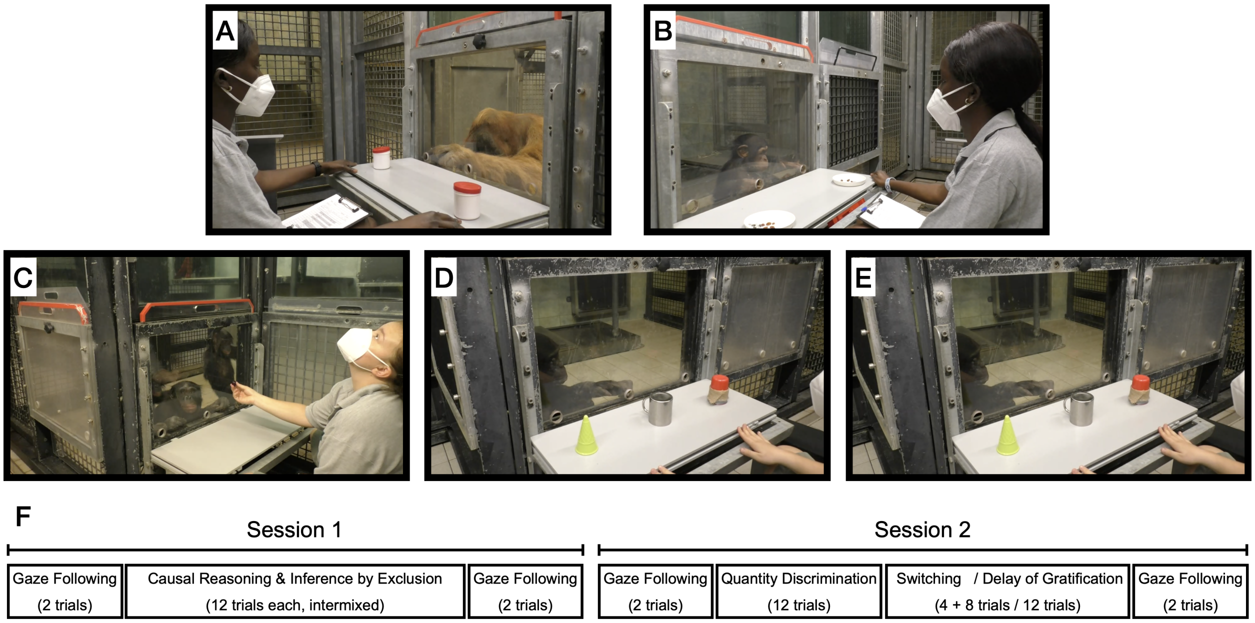

Figure S2: Setup used for the six tasks. A) Causal reasoning and inference by exclusion. B) Quantity discrimination. C) Gaze following. D) Switching. E) Delay of gratification.

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/`.

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 characteristics

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.

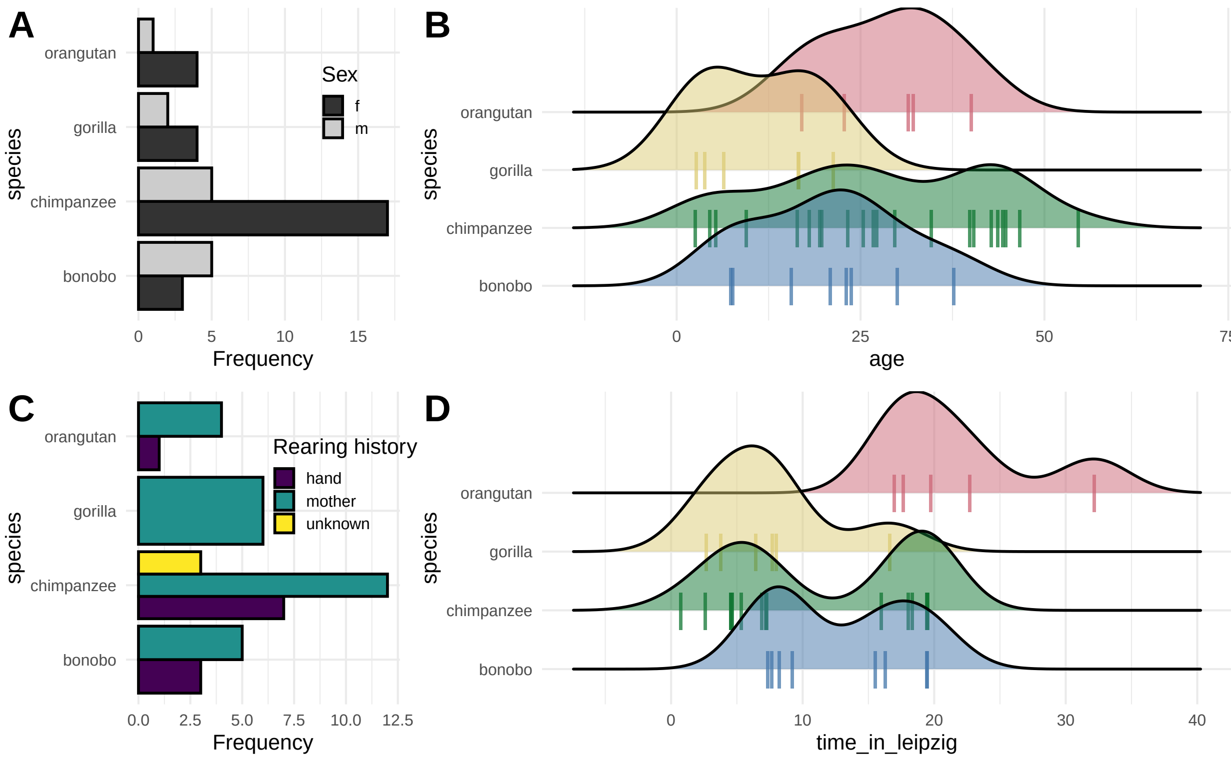

Figure S3: Stable individual characteristics. A) participant sex, B) age distribution by species, C) rearing history, D) time lived in leipzig by species.

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 S4A) 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 severe 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 workday (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 S4B visualizes the sociality measures for one group across the different time points.

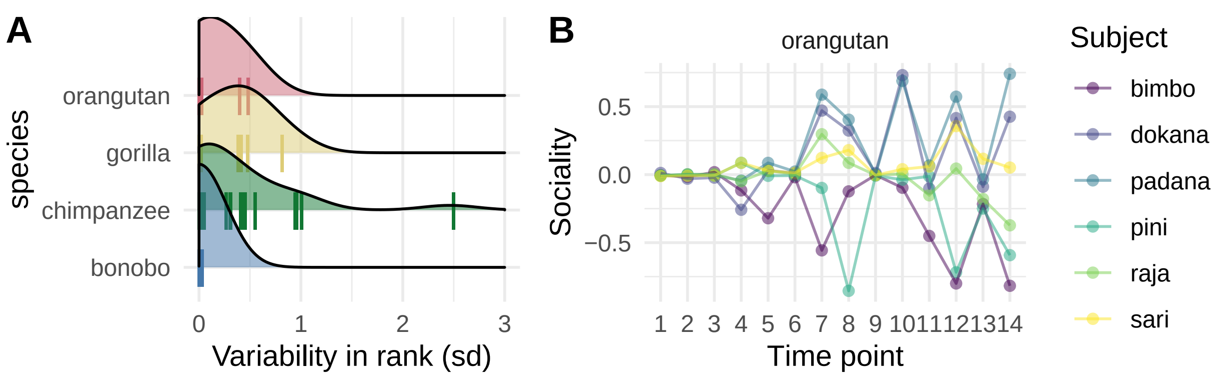

Figure S4: Variable individual characteristics. A) variability in rank (caretaker ratings) for each subject and species, B) sociality estimates for orangutans based on Multiple Membership Relations Model.

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 S5 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 caretakers to rate the severity of these events and averaged across them.

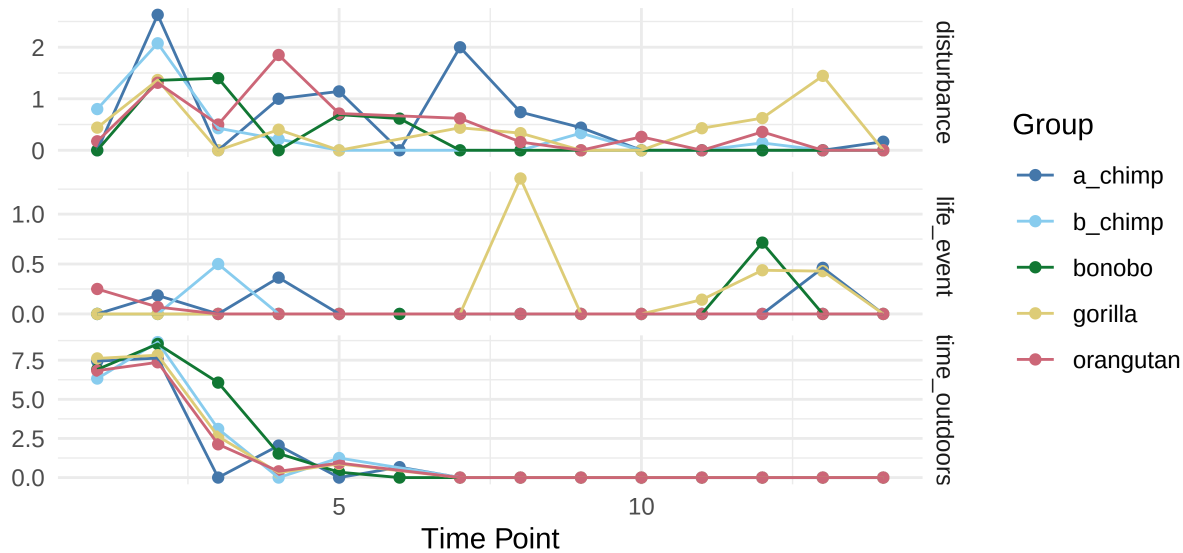

Figure S5: Variation in group life related measures across groups and time points.

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 recorded 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 in since the last time they were tested in that particular task. The experimenter took this information from the testing schedule.

Analytical framework

We have two overarching questions. On the one hand, we are interested in the stability and the reliability of the individual task as as well as the relations between them. We use structural equation modeling to address these questions. These models have been developed and are usually used with much larger sample sizes. Thus, we had to make a number of assumptions to be able to fit them to the kind of data that we have, which we mention in the text below.

On the other hand, we

Not analyse switching task

Two compoennts of the analysis and the different methods used for it

Structural equation modelling

Two parallel test halves were build, corresponding to sum scores of half of the trials of the same time point per task. Trials were alternately assigned to the first and the second test half. For tasks repeated 12 times per measurement time point this procedure resulted in two test halves assuming 7 possible values (0 to 6 correctly solved trials), for tasks repeated 8 times per measurement time point test halves could maximally assume 5 possible values (0 to 4 correctly solved trials). Not all categories were observed at all time points and categories had to be collapsed in some cases (see descriptions below). The two test halves served as indicators for a common latent construct per measurement time point, assuming parallel test halves (i.e., factor loadings set to 1 and assuming equal reliabilities). Due to only few observed categories, indicators were modeled as ordered categorical, using a probit link function. The models thereby correspond to Graded response models. For model parsimony, to improve estimation accuracy (see simulation studies) and in order to test for latent mean differences across time, thresholds for all indicators were set equal across time (resulting in the assumption of strict measurement invariance) as well as across test halves.

Models and coefficients

Latent State models Measurement equation for parcel i at time point t:

$$Y_{it} = S_t + \epsilon_{it} \tag{1}$$

At each time point t, a latent state variable S_t, underlying the two observed indicators Y_{1t} and Y_{it} is estimated. Latent state variables are allowed to freely correlate across time, with latent (measurement-error free) correlations serving as indirect indicators of stability across time. The model is depicted for 6 measurement time points in Figure S6.

Latent State Trait (LST) models Measurement equation for parcel i at time point t:

$$Y_{it} = T + S_t + \epsilon_{it} \tag{2}$$

where T is a stable latent trait variable, S_t captures time-specific deviations of the respective true score from the stable trait at time t, and ϵ_{it} is a measurement error variable, with $Var(\epsilon_{it}) = 1 \quad \forall i, t$ (probit parameterization; Graded response model). The model is depicted for 6 measurement time points in Figure S7.

Note that by assuming strong measurement invariance (i.e., loading parameters are set to 1 at all time points, residual variances are equal to 1 by definition of the Graded response model with a probit link, threshold parameters are set invariant across time points, variances of latent state residual factors are set invariant across time points), the specified LST model (without autoregressive effects) corresponds to a multilevel

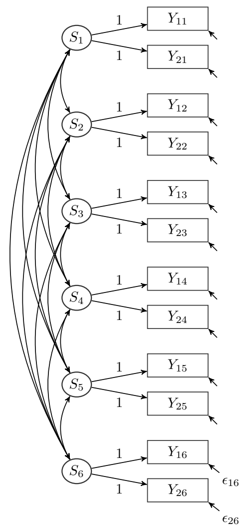

Figure S6: Latent State model for two indicators and six measurement time points.

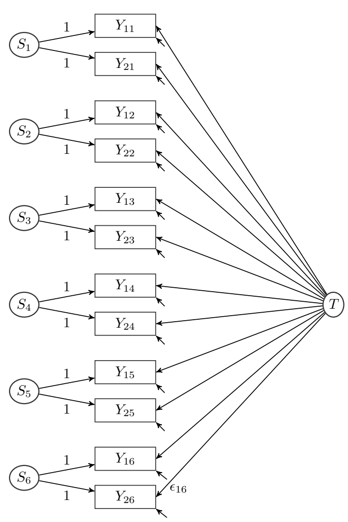

Figure S7: Latent State Trait model for two indicators and six measurement time points.

model with a latent trait factor at the between-level (person-level) and a latent state residual factor at the within-level (time-specific) level.

In order to test for possible mean changes across time, latent state models are estimated in a first step. LST models as single-level models are estimated to test whether measurement invariance assumptions across time can be reasonably assumed. Once measurement invariance can be established, the models can alternatively estimated as multilevel SEMs.

The following variance components can be computed for the presented LST model (without autoregressive effects).

Consistency Proportion of true variance (i.e., measurement-error free variance) that is due to true inter-individual stable trait differences.

$$Con(Y_{it}) = \frac{Var(T)}{Var(T) + Var(S_t)} \tag{3}$$

Occasion specificity Proportion of true variance (i.e., measurement-error free variance) that is due to true inter-individual differences in the state residual variables (i.e. occasion-specific variation).

$$OS(Y_{it}) = 1 - Con(Y_{it})) = \frac{Var(S_t)}{Var(T) + Var(S_t)} \tag{4}$$

As strong measurement invariance is assumed and $Var(S_t)$ is set equal across time, $OS(Y_{it})$ is constant across time as well as across item parcels i.

Latent State Trait (LST) models with autoregressive effects Model corresponds to model described in Eid et al. (2017). The model is depicted for 6 measurement time points in Figure S8.

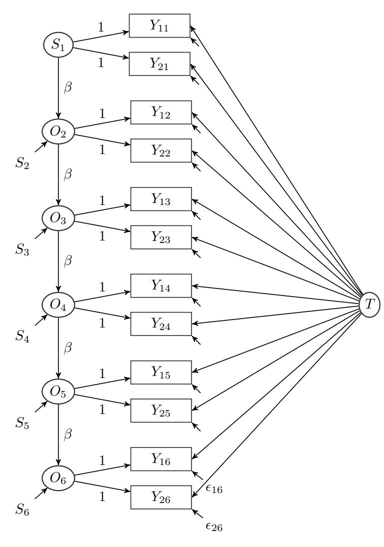

Figure S8: Latent State Trait model with autoregressive effects for two indicators and six measurement time points.

Measurement equation for parcel i at time point t:

$$Y_{it} = T + O_t + \epsilon_{it} \tag{5}$$

where T is a stable latent trait variable, O_t captures time-specific deviations of the respective true score from the stable trait at time t, and ϵ_{it} is a measurement error variable, with $Var(\epsilon_{it}) = 1 \quad \forall i, t$ (probit parameterization; Graded response model). O_t is assumed to follow an autoregressive process of order 1 across time (within subjects), that is:

$$
\begin{aligned}
O_t &= S_t \qquad t = 1 \\
O_t &= \beta O_{(t-1)} + S_t \qquad t > 1
\end{aligned}
$$

where the latent state residual variables S_t capture true occasion-specific inter-individual differences that can not be explained by states at previous measurement time points.

The following variance coefficients can be computed.

Consistency Proportion of true variance (i.e., measurement-error free variance) that is due to true inter-individual stable trait differences.

$$Con(Y_{it}) = \frac{Var(T)}{Var(T) + \beta^2 Var(O_{(t-1)}) + Var(S_t)} \tag{6}$$

Occasion specificity Proportion of true variance (i.e., measurement-error free variance) that is due to true inter-individual differences in the state residual variables, that is occasion-specific variation that is not explained by the autoregressive process.

$$OS(Y_{it}) = \frac{Var(S_t)}{Var(T) + \beta^2 Var(O_{(t-1)}) + Var(S_t)} \tag{7}$$

As the proportion of variance explained by the autoregressive process stabilizes over time, all coefficients have converged to a relatively stable value at $t = 14$, indicating the long-term proportions of variance that are to be expected.

Predictability Proportion of true variance that is explained by carry-over effects from previous measurement time points.

$$Pred(Y_{it}) = \frac{\beta^2 Var(O_{(t-1)}}{Var(T) + \beta^2 Var(O_{(t-1)}) + Var(S_t)} \tag{8}$$

Estimation

Models were estimated with MPlus version 8.4, using Bayesian Markov-Chain Monte-Carlo sampling, with the Mplus default priors (also see simulation studies). Using inverse gamma priors [IG(0.001, 0.001); see simulation study] for LST models did not substantially change the parameter estimates. Therefore, only the results with respect to the MPlus default priors are reported. We used two chains with a minimum of 10,000 iterations per chain, with a thinning of 10 (corresponds to a minimum of 100,000 drawn samples per chain of which every 10th is used for the construction of the posterior distribution). The first half of each chain is discarded as burn-in. Convergence was assumed and estimation stopped when the Potential Scale Reduction (PSR) factor well below a threshold of 1.01 for the first time after the minimum number of iterations was reached.

Projection predictive inference

The predictive projection approach developed by Piironen, Paasiniemi, and Vehtari (2018) was used to select a minimal subset of covariates to find a predictive model for cognitive performance. Projective selection can be viewed as a two-step process. The first step revolves around building the best predictive model possible, called the reference model. The reference model is a Bayesian multilevel regression model (repeated measurements nested in apes), including all covariates recorded in this study. In the second step, the goal is to replace the posterior distribution of the reference model with a simpler distribution. This is achieved via a forward stepwise addition of covariates that decrease the Kullback-Leibler divergence from the reference model to the projected model. The result is a list containing the best model for each number of covariates. The final model is selected by inspecting the mean log predictive density. The projected model with the smallest number of covariates shows similar predictive performance as the reference model is chosen.

The advantages of predictive projection are that it provides an excellent tradeoff between model complexity and accuracy Piironen and Vehtari (2017). It has also been shown that predictive projection is useful when identifying **all relevant** covariates is of importance.

The predictive projection technique is implemented in the R package 'projpred' "Projpred: Projection Predictive Feature Selection" (n.d.).

We built a reference model for each task, resulting in four reference models and four rankings of relevant predictors.

Results

Phase 1

Stability

Reliability

Relations between tasks

Predictability

Summary

Appendix

SEM Simulations

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.

Eid, Michael, Jana Holtmann, Philip Santangelo, and Ulrich Ebner-Priemer. 2017. "On the Definition of Latent-State-Trait Models with Autoregressive Effects: Insights from LST-r Theory." *European Journal of Psychological Assessment* 33 (4): 285.

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.

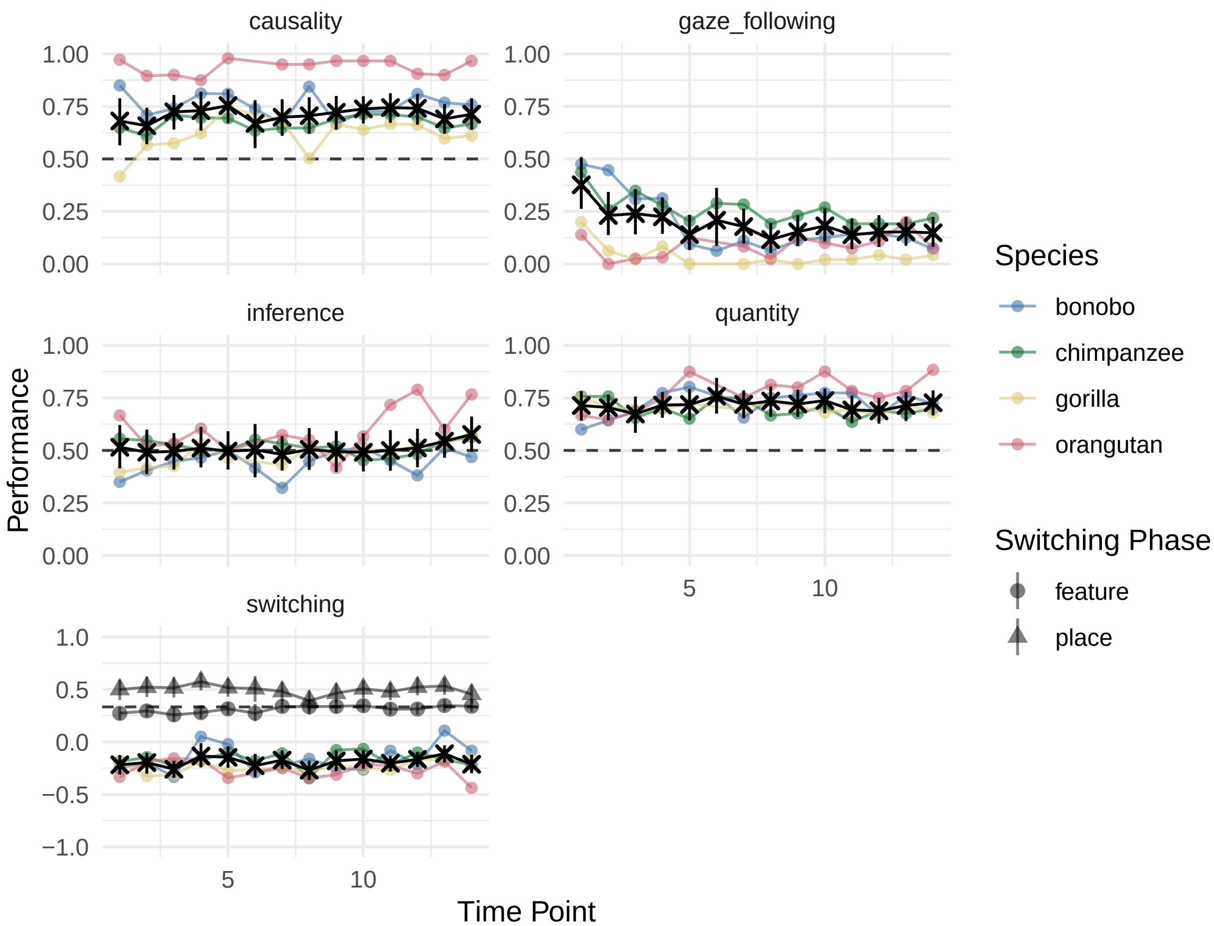

Figure S9: Results from the five cognitive tasks across time points. Black crosses show mean performance at each time point across species (with 95% CI). Colored dots show mean performance by species. Dashed line shows the chance level whenever applicable. The panel for switching includes triangles and dots showing the mean performance in the two phases from which the overall performance score was computed.

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.

Piironen, Juho, Markus Paasiniemi, and Aki Vehtari. 2018. "Projective Inference in High-Dimensional Problems: Prediction and Feature Selection." *arXiv Preprint arXiv:1810.02406*.

Piironen, Juho, and Aki Vehtari. 2017. "Comparison of Bayesian Predictive Methods for Model Selection." *Statistics and Computing* 27 (3): 711–35.

"Projpred: Projection Predictive Feature Selection." n.d. https://mc-stan.org/projpred.

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