Contents

Overview

This document supplements the paper "Individual differences in great ape cognition across time and domains: stability, structure, and predictability". Some text passages and figures are the same in the supplementary material and the main paper. This redundancy is intended and ensures that the supplementary material is self-contained and readable.

This document is structured as follows: we first describe the sample. Next we describe the general setup and the experimental tasks. In the section data collection, we lay out the time line of data collection. Next, we give an overview of the predictor variables we recorded in addition to the experimental data.

We then move on to our analytical framework. First we describe the Structural Equation models that were used to investigate robustness, stability and reliability of cognitive performance. Next, we describe the Projection Predictive Inference analysis which we used to test the importance of the predictor variables.

In the results we first report on the robustness, stability and reliability of each task. Next, we investigate relations between trait variables computed for the different tasks. Here we also include the tasks described in Bohn et al. (2023). Finally, we report how the different predictors related to performance in the different tasks.

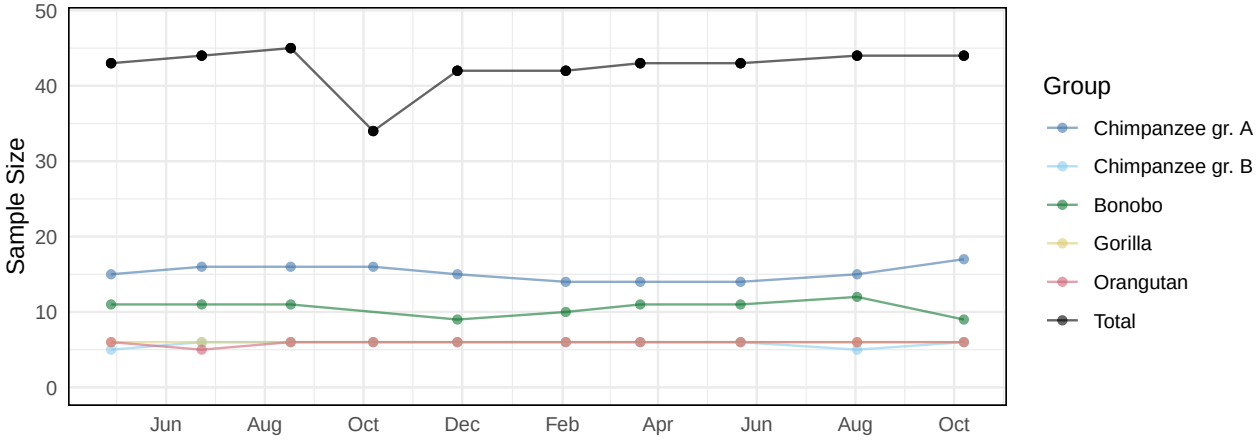
Methods

Participants

A total of 48 great apes participated at least in one tasks at one time point. This included 12 Bonobos (4 females, age 3.6 to 40.7), 24 Chimpanzees (17 females, age 3.8 to 57.8), 6 Gorillas (4 females, age 4.4 to 24.4), and 6 Orangutans (5 females, age 4.7 to 43.1). The sample size at the different time points ranged from 34 to 45 for the different species. Figure @ref(fig:sample) 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 logistical constraints (e.g. construction works). All apes participated in cognitive research on a regular basis. Many of them had experience with the tasks we used in the current study (see predictor variables).

Apes were housed at the Wolfgang Köhler Primate Research Center located in Zoo Leipzig, Germany. They lived in groups, with one bonobo, gorilla and orangutan group, respectively, and two chimpanzee groups

(group A and B). 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.



Supplementary Figure 1: Sample size by species across the different time points. Time point specific predictor variables were collected during the time between two time points 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 mesh or 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 Supplementary Figure @ref(fig:setup)).

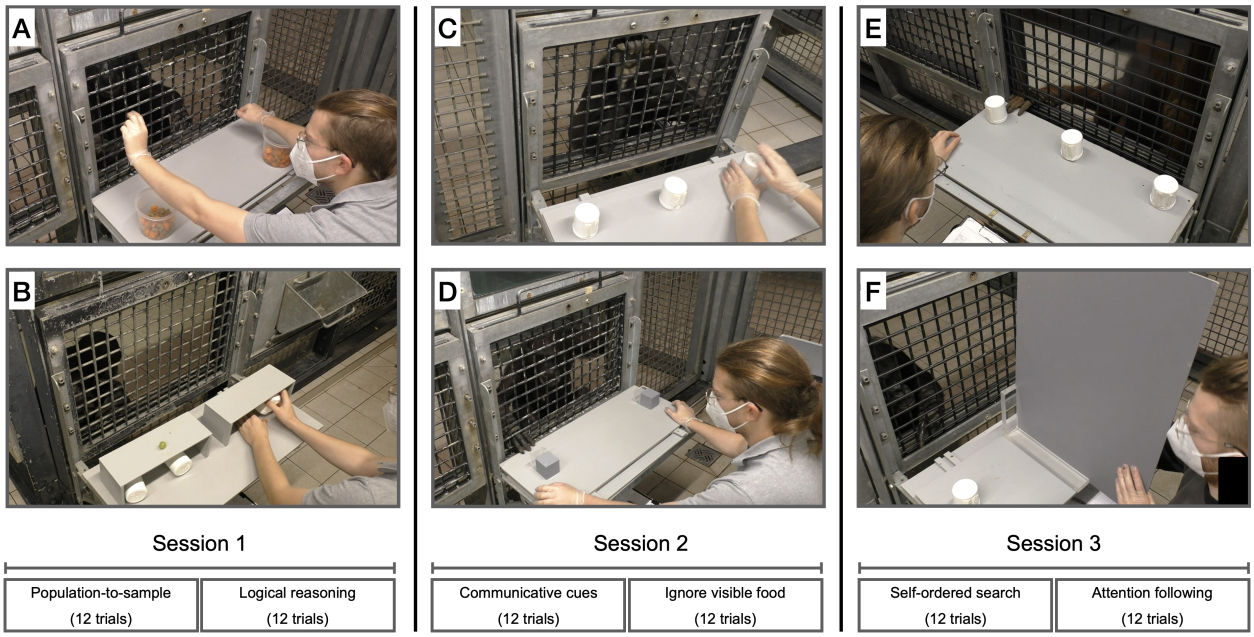
Tasks

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, cognitively less demanding ways to solve the tasks. 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/`.

Attention-following

The Attention-following task was loosely modeled after Kaminski, Call, & Tomasello (2004). The setup consisted of two identical cups placed on the sliding table and a large opaque screen that was longer than the width of the sliding table (Supplementary Figure @ref(fig:setup)F). The experimenter placed both cups on the table and showed the ape that they were empty. Then, the experimenter baited both cups in view of the ape and placed the opaque screen in the center between the two cups, perpendicular to the mesh. Next, the experimenter moved to one side and looked at the cup in front of them. Then, the experimenter pushed the sliding table forward and the ape was allowed to choose one of the cups by pointing at it. If the ape chose the cup that the experimenter was looking at, they received the food item. If they choose the other cup, they did not. Because the experimenter could not see if the ape chose the other cup, the trial was terminated



Supplementary Figure 2: Setup used for the six tasks. A) population-to-sample, B) logical-reasoning, C) communicative-cues, D) ignore-visible-food, E) self-ordered-search and F) attention-following. Text at the bottom shows order of task presentation and trial numbers

if the ape was at the mesh (in a position to make a choice) and did not choose the cup the experimenter was looking at within 3s. We coded whether the ape chose the side the experimenter was looking at (correct choice) or not. Apes received twelve trials. The side at which the experimenter looked was counterbalanced with same number of looks to each side and looks to the same side not more than two times in a row.

We assume that apes follow the experimenters focus of attention to determine whether or not their request could be seen and thus be successful.

Communicative-cues

This task was modeled after Schmid, Karg, Perner, & Tomasello (2017). Three identical cups were placed equidistantly on a sliding table directly in front of the ape (Supplementary Figure @ref(fig:setup)C). In the beginning of a trial, the experimenter showed the ape that all cups are empty. After placing an occluder between the subject and the cups, the experimenter held up one food item and moved it behind the occluder, visiting all three cups but baiting only one. Next, the occluder was lifted and E looked at the ape (ostensive cue), called the name, and looked at one of the cups, while holding on to it with one hand and tapping it with the other (continuous looking, 3 times tapping). Finally, the experimenter pushed the sliding table forward for the ape to make a choice. If the ape chose the baited cup, they received the reward – if not, not. We coded as correct choice if the ape chose the indicated cup. Apes received twelve trials. The location of the indicated cup was counterbalanced, with each cup being the target equally often and the same target not more than two times in a row.

We assume that apes use the experimenter's communicative cues to determine where the food is hidden.

Ignore-visible-food

The task was modeled after Völter, Tinklenberg, Call, & Seed (2022). The task involved two opaque cups with an additional, sealed but transparent, compartment attached to the front of each cup (facing the ape). For one cup, the compartment contained a preferred food item that was clearly visible, for the other cup, the compartment was empty (Supplementary Figure @ref(fig:setup)D). In the beginning of the trial, the two cups

were placed upside down on the sliding table so that the ape could see that the opaque compartments of both cups were empty. Next, the experimenter baited one of the cups in full view of the subject. In non-conflict trials, the baited cup was the cup with the food item in the transparent compartment. In conflict trials, the baited cup was the cup with the empty compartment. After baiting the experimenter pushed the sliding table forwards and the ape could chose by pointing. If the baited cup was chosen, the ape received the food. Apes received 14 trials, twelve conflict trials and two non-conflict trials (1st and 8th trial). Only conflict trials were analyzed. The location of the cup with the baited compartment was counterbalanced, with the cup not being in the same location more than twice in a row.

We assume that apes need to inhibit selecting the visible food item and instead use their short-term memory to remember where the food was hidden.

Logical-reasoning

The task was modeled after Hanus & Call (2014). Three identical cups were presented side-by-side on a sliding table, with the cup in the middle sometimes positioned closer to the left cup and sometimes closer to the right. (Supplementary Figure @ref(fig:setup)B). Two half-open boxes served as occluders to block the ape's view when shuffling the cups. Each trial started by showing the ape that all three cups (one on one side of the table, two on the other) were empty. After placing the occluders over both sides of the table, the experimenter put one piece of food on top of each occluder. Next, the experimenter hid each piece of food under the cup(s) behind the occluders. In case of the occluder with the two cups, the food was randomly placed under one of the two cups while both cups were visited and even shuffled. Finally, both occluders were lifted and the table pushed forwards, allowing the ape to choose one of the three cups, from which they then received the content. We coded whether the ape chose the certain cup (i.e. the cup from the side of the table with only one cup). Apes received 12 trials. The side with one cup was counterbalanced, with the same constellation appearing not more than two times in a row on the same side.

We assume that apes would infer that the cup from the tray with only one cup certainly contains food while the other cups contain food only in 50% of cases.

Population-to-sample

The task was modeled after Eckert, Call, Hermes, Herrmann, & Rakoczy (2018). During the test, apes saw two transparent buckets filled with pellets and carrot pieces (the carrot pieces had roughly the same size and shape as the pellets). Each bucket contained 80 food items. The distribution of pellets to carrot pieces was 4:1 in bucket A, and 1:4 in bucket B. Pellets are preferred food items in comparison to carrots. The experimenter placed both buckets on a table, one left, one right (Supplementary Figure @ref(fig:setup)A). In the beginning of a trial, the experimenter picked up the bucket on the right side, tilted it forward so the ape could see inside, placed it back on the table and turned it around 360°. The same procedure was repeated with the other bucket. Next, the experimenter looked at the ceiling, inserted each hand in the bucket in front of it and drew one item from the bucket without the ape seeing which type (E picked always of the majority type). The food items remained hidden in the experimenter's fists. Next, the experimenter extended the arms (in parallel) towards the ape who was then allowed to make a choice by pointing to one of the fists. The ape received the chosen sample. In half of the trials, the experimenter crossed arms when moving the fists towards the ape to ensure that the apes made a choice between samples and not just chose the side where the favorable population was still visible. In between trials, the buckets were refilled to restore the original distributions. Apes received twelve trials. We coded whether the ape chose the sample from the population with the higher number of high quality food items. The location of the buckets (left and right) was counterbalanced, with the buckets in the same location no more than two times in a row. The crossing of the hands was also counterbalanced with no more than two crossings in a row.

We assume that apes reasoned about the probability of the sample being a high quality item based on observing the ratio in the population.

Self-ordered-search

The task was modeled after Völter, Mundry, Call, & Seed (2019; Diamond, Prevor, Callender, & Druin, 1997; see also Petrides, 1995). Three identical cups were placed equidistantly on a sliding table directly in front of the ape (Supplementary Figure @ref(fig:setup)E). The experimenter baited all three cups in full view of the ape. Next, the experimenter pushed the sliding table forwards for the ape to choose one of the cups by pointing. After the choice, the table was pulled back and the ape received the food. After a 3s pause, the table was pushed forward again for a second choice. This procedure was repeated for a third choice. If the ape chose a baited cup, they received the food, if not, not. We coded the number of times the ape chose and empty cup (i.e. chose a cup they already chose before). Please note that this outcome variable differed from the other tasks in two ways: first, possible values were 0, 1, and 2 (instead of just 0 and 1) and second, a lower score indicated better performance. Apes received twelve trials. No counterbalancing was needed.

We assume that apes use their working memory abilities to remember where they had already searched and which cups still contained food.

Data collection

One time point meant running all tasks with all participants. Within each time point, the tasks were organized in three sessions (see Supplementary Figure @ref(fig:setup)). Session 1 included the population-to-sample and logical-reasoning tasks, session 2 the communicative-cues and ignore-visible-food tasks and session 3 the self-ordered-search and attention-following tasks.

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/`. This exact procedure was repeated at each time point so that the results would be comparable across participants and time points. The three sessions were usually spread out across three adjacent days. For the larger chimpanzee group, they were spread out across six days, but so that each individual completed the tasks within thre days (a.i. 50% of the group where tester in the first three days, the other half in the last three days).

Data collection started on April 28th, 2022, lasted until October 7th, 2023 and included 10 time points. The interval between two time points was planned to be eight weeks. However, it was not always possible to follow this schedule so that some intervals were longer or shorter. Supplementary Figure @ref(fig:sample) visualizes the intervals between time points. All the data was collected by the same experimenter and where live-coded within the session. The experimenter was alone in the test room most of the time; on rare occasions there was a second person present to observe (e.g., for training purposes or internships).

Predictors

In addition to the data from the cognitive tasks, we collected data for a range of predictor variables. The goal 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 and thus with individual, time point and session (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.

Stable individual characteristics

These predictors are stable individual differences. As a source, we used the ape handbook at Zoo Leipzig. Supplementary Figure @ref(fig:demo) gives an overview of the distribution of the different characteristics in

the sample.

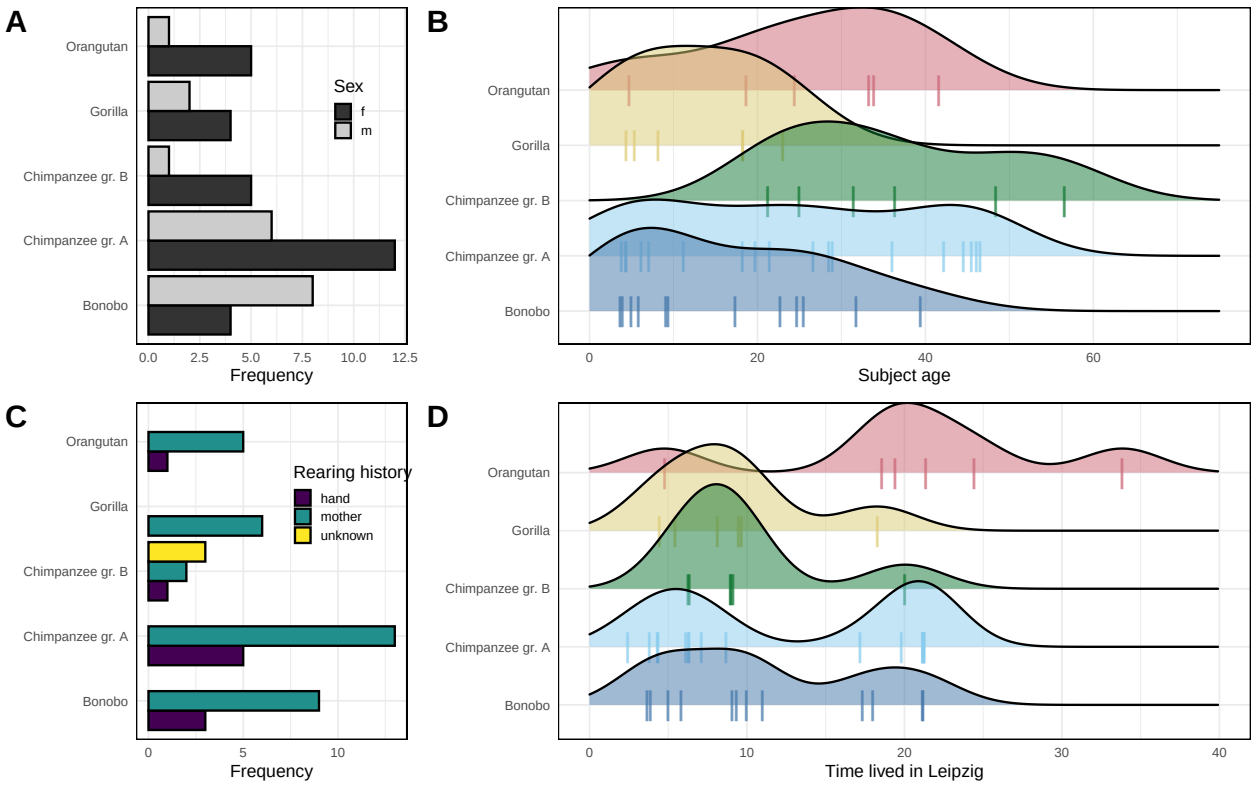
Group Group the individual belonged to. Groups were composed of individuals of the same species but because there were two chimpanzee groups (A-chimpanzees and B-chimpanzees), group and species are not equivalent. Variable name in model: `Group`.

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. Variable name in model: `Age`.

Sex Participant's biological sex. Variable name in model: `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. Variable name in model: `Rearing history`.

Time spent in research Absolute time the individual has lived in Leipzig Zoo. All apes living in Leipzig are involved in behavioral research to a certain degree. Some groups, like the A-chimpanzee group are involved in more studies compare to e.g., the Gorillas. These more fine-grained differences are captured in a different variable described below (`Study participation (time point)`). Thus, we take this measure to be a rough proxy of how much experience an individual has had with cognitive research in general. Variable name in model: `Time spent in research`.



Supplementary Figure 3: Stable individual characteristics. A) participant sex, B) age distribution by species, C) rearing history, D) time lived in leipzig by group.

Variable individual characteristics

These predictors varied by participant and time point.

Rank We asked caretakers to order individuals within a given group according to their rank. Ties were allowed. This was done at each time point. An individual's rank was mostly stable (see Supplementary Figure @ref(fig:socrel)A) across time points, however, there was some variation. Variable name in model: `Rank`.

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. Variable name in model: `Sickness severity`.

Sociality We conducted proximity scans for all groups in the early afternoon on every workday (Monday to Friday)[*Nico: check if this is roughly correct*]. For each individual, we recorded which individuals were within arms reach. Research assistants used a tablet to record their observations with the behavioral coding software ZooMonitor (Wark et al., 2019). Given the variable intervals between time points and constraints due to the availability of personnel, we did not have the same number of scans for each time point and species. On average, there were 38.69 scans per subject and time point (range: 14 - 68).

To derive individual specific estimates of sociality for each time point, we fit a variant of a Social Relations Model (Snijders & 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 models 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, Whalen, Koster, & Perry (2021) suggested to speak of a Multiple Membership Relations Model (see also Leckie, 2019) in such a context, which simply estimates how 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.

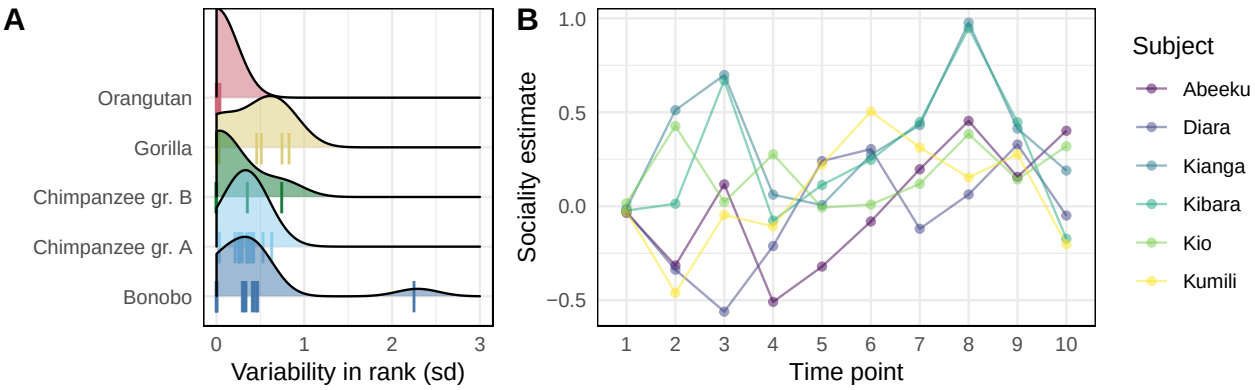
For each individual and time point, we extracted the sociality estimates and used them to predict cognitive performance in the different tasks for that time point. Supplementary Figure @ref(fig:socrel)B visualizes the sociality measures for one group across the different time points. Variable name in model: `Sociality`.

Group life

These 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. Supplementary Figure @ref(fig:glife) 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 instead of the indoor enclosure or the sleeping rooms. To compute the predictor, we averaged across these values for each time point and group. Variable name in model: `Time spent outdoors`.

Disturbances The animal caretakers also noted if there were any unusual disturbances for a particular group. Examples 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. Variable name in model: `Disturbance`.



Supplementary Figure 4: Variable individual characteristics. A) variability in rank (caretaker ratings) for each subject and species, B) sociality estimates for gorillas based on Multiple Membership Relations Model.

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 on a scale from 1 to 7 and averaged across them. Variable name in model: `Life event`.

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. Supplementary Figure @ref(fig:gtest) visualizes the different variables across time points.

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. Variable name in model: `Observer present`.

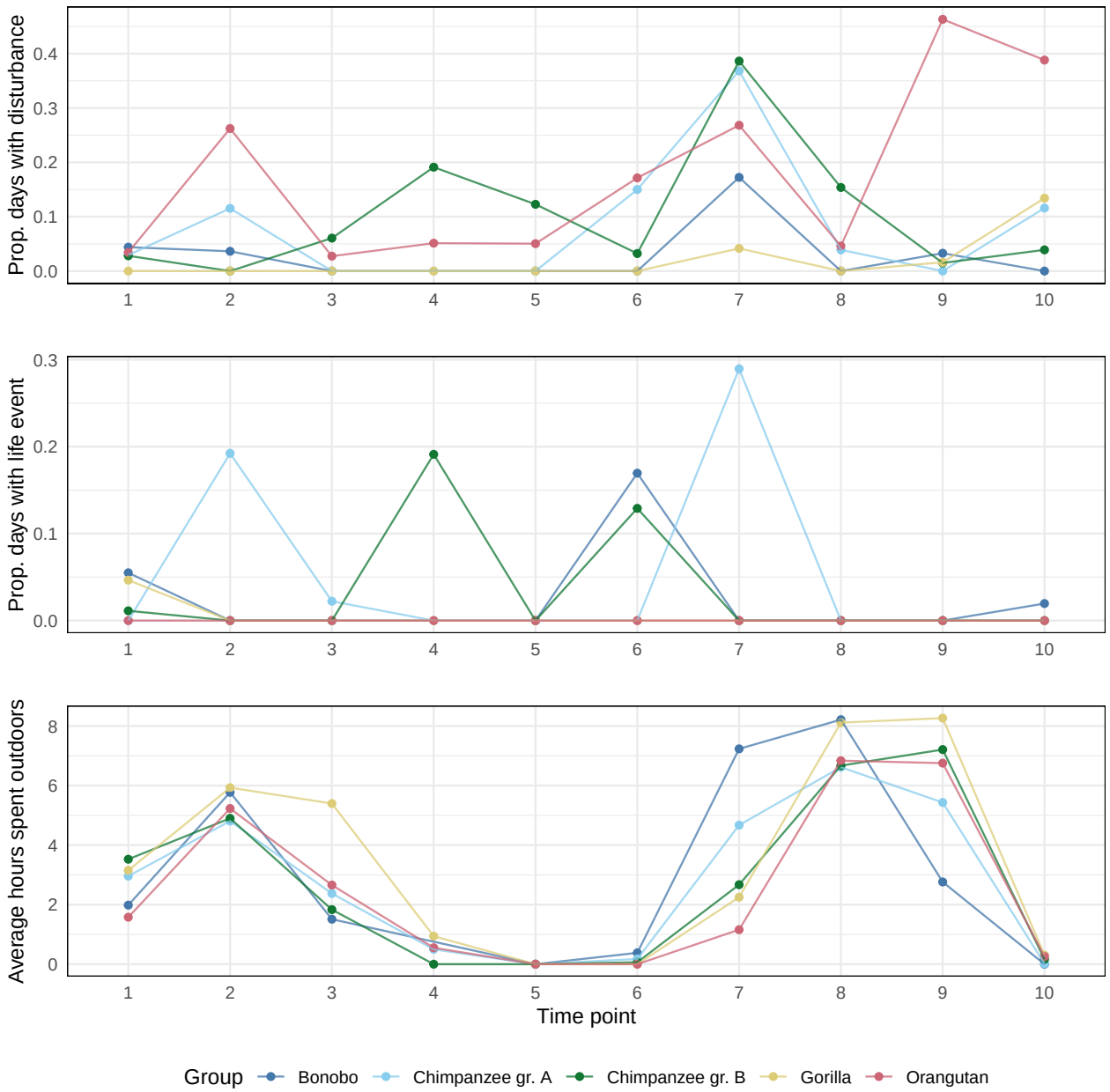
Study on same day This predictor recorded whether or not the participant had already participated in a different study on the same day. The experimenter took this information from the testing schedule. Variable name in model: `Study participation (day)`.

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. Variable name in model: `Study participation (time point)`.

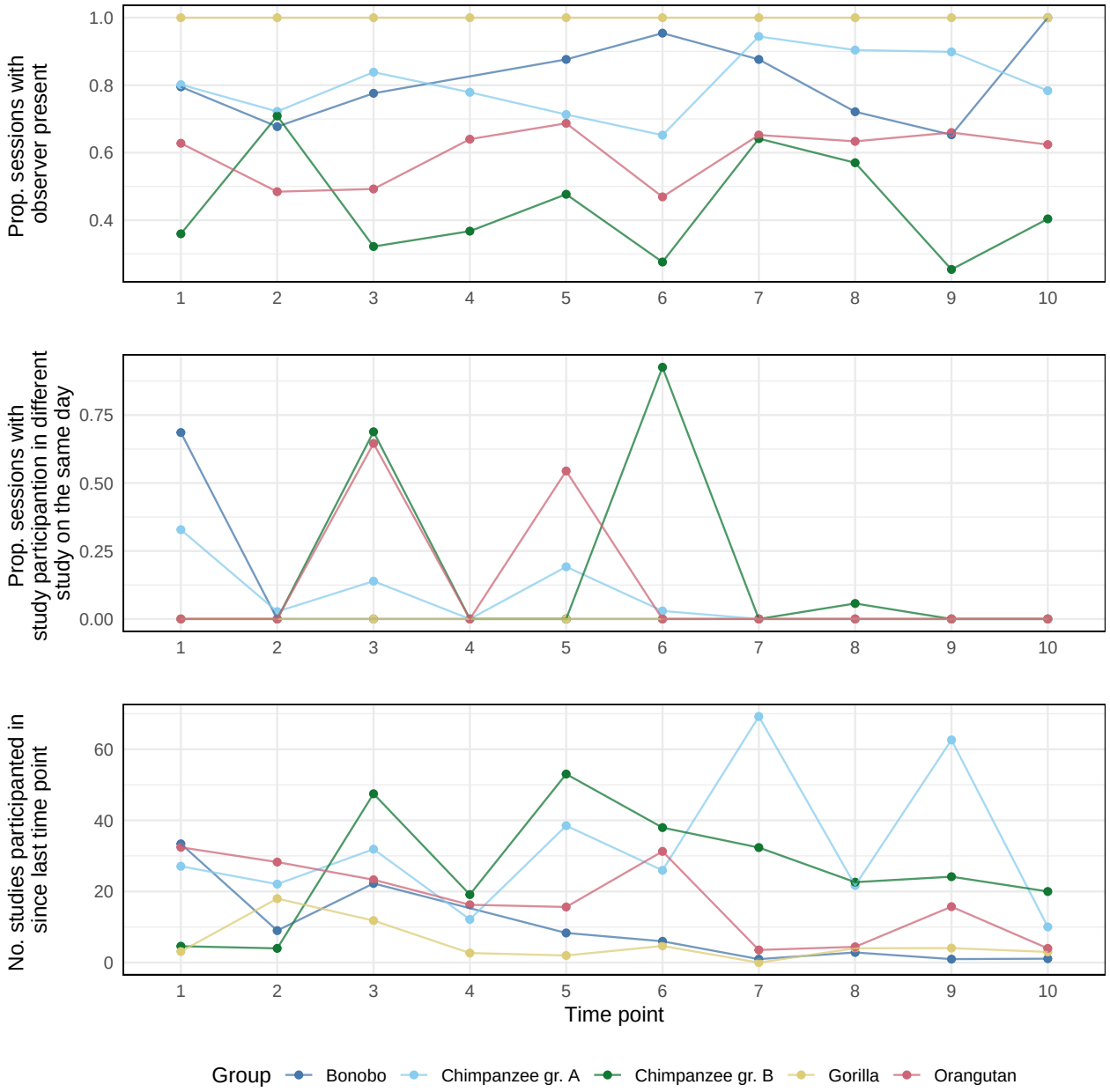
Analytical framework

We had two overarching questions. On the one hand, we were interested in the cognitive measures and the relations between them. That is, we asked how robust performance in a given task was on a task-level, how stable individual differences were and how reliable the measures were. We also investigated relations between the different tasks. We used *Structural Equation Modeling* (SEM) (Bollen, 1989; Hoyle, 2012) to address these questions. SEMs usually require larger sample sizes than available in the present study. In the Supplementary Material of Bohn et al. (2023) we reported a small simulation study which showed that parameters in the employed SEMs could be accurately estimated using Bayesian estimation techniques given our available sample sizes under reasonable model restrictions. We lay out the restrictive assumptions we imposed on the parameters in the text below.

Our second question was, which predictors explain variability in cognitive performance. Here we wanted to see which of the predictors we recorded were most important to predict performance over time. This



Supplementary Figure 5: Variation in group life related measures across groups and time points.



Supplementary Figure 6: Variation in testing arrangenments across groups and time points. Top: Proportion of individuals that had an observer present while being tested. Middle: Proportion of individuals who participated in a different study on the same day. Bottom: Average number of studies individuals participated in between time points.

is a variable selection problem (selecting a subset of variables from a larger pool) and we used *Projection Predictive Inference* for this (Piironen, Paasiniemi, & Vehtari, 2020).

Structural equation modeling

In the present analyses we were interested in estimating the stability of performances in a given task across time as well as the association between performances across different tasks. To separate components of random fluctuation (measurement error) from systematic differences in performance across time, we used Structural equation models (SEM). SEMs can be used to model relations between latent variables (constructs) which are estimated based on several observed variables.

[Jana: following paragraph to be updated to include LST model with variable means] We used latent state-trait models to separate traits (stable over time) from state residuals (time varying). In the present context, one can think of a trait as a stable psychological ability (e.g. ability to make causal inferences) and state residuals as time-specific deviations from these traits due to variable psychological conditions (e.g. variations in performance due to being attentive or inattentive). Variation in performance on a given time point can then be partitioned into variance due to the trait, variance due to the situation or individual-situation interactions, and measurement error. Because the latent variables are estimated on multiple indicators (here: test halves), they are assumed to be measurement-error free (Geiser, 2020; Steyer, Ferring, & Schmitt, 1992; Steyer, Mayer, Geiser, & Cole, 2015). Next we describe the model construction process in more detail.

At each time point, we observed 12 identical trials per individual per task. We modeled sum scores of the repeated trials (instead of latent ability factors), given that each trial per task was an identical repetition of the same task.

To separate reliable from unreliable variance components and obtain reliability estimates of the resulting sum score variables, we build two sum score variables per task per time point. That is, for each task, two parallel test halves were build, corresponding to performance 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 all tasks except self-ordered-search this procedure resulted in two test halves with 7 possible values (0 to 6 correctly solved trials). For self-ordered-search, because one could search in the same location twice on each trial, test halves could maximally assume 13 possible values (0 to 12 errors).

We interpreted reliability estimates in the following way: low < 0.5, moderate $= 0.6$, acceptable $= .7$, good $= .8$ and high $= .9$. Please note that these estimates are for test-halves and the reliability of the full test would be higher.

The two test halves served as indicators for a common latent construct per time point, assuming parallel test halves (i.e., factor loadings set to 1 and assuming equal reliability). Due to only few different observed values and skewed distributions of the sum score variables, indicators were modeled as ordered categorical variables, using a probit link function. The models thereby correspond to normal-ogive graded response models (Samejima, 1969, 1996). That is, the models assume a continuous latent ability underlying the discrete responses, with an increasing probability of more correctly solved trials with increasing ability.

For model parsimony, to improve estimation accuracy, and in order to test for latent mean differences across time, we assume strict factorial (or measurement) invariance across time (Meredith, 1993; Millsap & Yun-Tein, 2004). That is, in each model (task), loading parameters are set to 1 at all time points, residual variances are equal to 1 (by definition of the graded response model as detailed below), and threshold parameters (see below for details) are set invariant across time points. In other words, we assume that the indicators (test halves) measure the latent variable in an equivalent and stable manner over time.

Models and coefficients

For each task, we constructed three different models which increased in complexity. We started with a latent state (LS) model, which estimates a latent state for each time point based on the two test halves. Robustness of task-level performance can be assessed by comparing latent state means across time points. Stability of individual differences can be assessed by correlating latent state variables across different time points.

Second, we fit a latent state-trait (LST) model. In LST models, true inter-individual differences are decomposed into a latent trait variable and time-specific deviations of the true score from the latent trait (state residual variable). In the following LST models, we assumed stable latent trait variables across time (no trait change). The model allowed us to partition the true variance in performance into stable (trait) and variable (state residual) components. Assuming a stable latent trait variable, the LST model is more restrictive than the LS model with respect to the implied covariance matrix, as correlations between true scores are not freely estimated across time points but assumed to be the same for different time lags.

[Jana: quick explanation of this model] Third, we fit LST models with varying means ...

The following sections give a mathematical description of the different models and the parameters in them.

Latent state models In the following we chose a factor analytical representation of the graded response model, that is, we present the models as factor models for ordinal data. Thereby we assume that the observed categorical variables Y_{it} for test half i at time point t result from a categorization of unobserved continuous latent variables Y_{it}^* which underlie the observed categorical variables. For observed variables that take on k_{it} different ordered values out of the set of possible categories $S_{it} = 0, ..., k_{it} - 1$ the relation between Y_{it} and Y_{it}^* is described by:

$$Y_{it} = \begin{cases} 0 & for \quad Y_{it}^* \leq \kappa_{1it} \\ s & for \quad \kappa_{sit} < Y_{it}^* \leq \kappa_{(s+1)it} \qquad with \;\; 0 < s < k_{it} - 1 \\ k_{it} - 1 & for \quad \kappa_{(k_{it}-1)it} < Y_{it}^* \end{cases} \tag{1}$$

where κ_{sit} denote threshold parameters (B. Muthén, 1984).

The graded response model assumes that the different categories of responses (in our case the number of correct trials per test half) form an ordered scale. Which category an individual scores depends on their latent ability. Because the latent variable is continuous but the response is discrete, there are thresholds on the latent ability that mark the transition between response categories. The threshold parameters κ_{sit} correspond to the level of the latent ability necessary to respond in category s or higher with 0.50 probability.

In latent state models, these continuous latent variables Y_{it}^* are decomposed into a latent state variable S_t and a measurement error variable ϵ_{it} (see, for instance Eid & Kutscher, 2014):

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

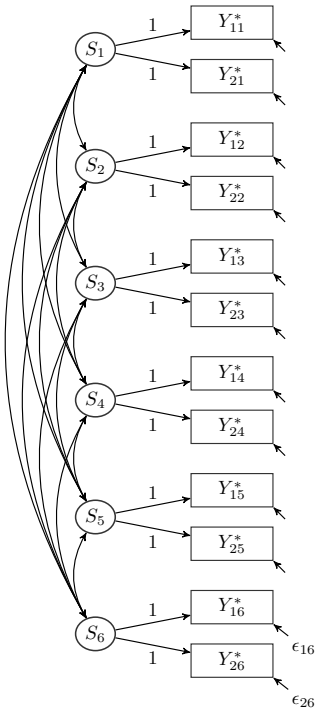
with $\epsilon_{it} \sim N(0, 1) \; \forall \; i, t$ (probit parameterization; normal-ogive graded response model). See Takane & De Leeuw (1987) for the equivalence of the normal-ogive graded response model and the factor model with ordinal indicators.

At each time point t, the two latent variables Y_{1t}^* and Y_{2t}^* are assumed to capture a common latent state variable S_t. 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 six measurement time points in Supplementary Figure @ref(fig:lsgraph).

To test for possible mean changes of ability across time, the means of the latent state variables are freely estimated (assuming invariance of the threshold parameters κ_{sit} across time).

As an estimate of reliability, the proportion of true score variance relative to the total variance of the continuous latent variables Y_{it}^* is computed:

$$Rel(Y_{it}^*) = \frac{Var(S_t)}{Var(S_t) + Var(\epsilon_{it})} = \frac{Var(S_t)}{Var(S_t) + 1} \tag{3}$$



Supplementary Figure 7: Latent State model for two indicators and six measurement time points.

Latent state-trait (LST) models In LST models, the latent state variables S_{it} are further decomposed into a latent trait variable T_{it} and a latent state residual variable ζ_{it}. The latent trait variables T_{it} are time-specific dispositions, that is, trait scores capture the expected value of the latent state (i.e., true score) variable for an individual at time t across all possible situations the individual might experience at time t (Eid, Holtmann, Santangelo, & Ebner-Priemer, 2017; Steyer et al., 2015). The state residual variables ζ_{it} capture the deviation of a momentary state from the time-specific disposition T_{it}. In the following models, we assume that latent traits are stable across time. Additionally assuming common latent trait and state residual variables across the two test halves, results in the following measurement equation for parcel i at time point t:

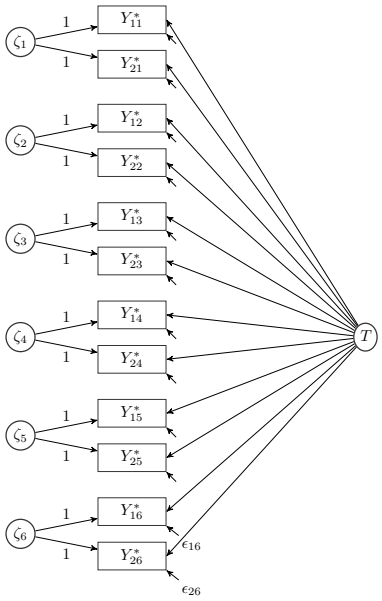
$$Y^*_{it} = T + \zeta_t + \epsilon_{it} \tag{4}$$

Here, T is a stable (time-invariant) latent trait variable, capturing stable inter-individual differences between individuals. The state residual variable ζ_t captures time-specific deviations of the respective true score from the trait variable at time t, and thereby captures deviations from the trait due to situation or person-situation interaction effects. ϵ_{it} denotes a measurement error variable, with $\epsilon_{it} \sim N(0, 1)\ \forall\ i, t$. The model is depicted for 6 measurement time points in Supplementary Figure @ref(fig:lstgraph).

As noted above, we assume strict factorial (measurement) invariance. Additionally, for reasons of parsimony, we assume that the variances of the state residual variances are invariant across time. As a consequence, the specified LST model corresponds to a multilevel 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.

The following variance components can be computed for the presented LST model.

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



Supplementary Figure 8: Latent State-Trait model for two indicators and six measurement time points. All factor loadings of the latent trait factor T are fixed to 1 (not displayed in the figure)

$$Con(Y^*_{it}) = \frac{Var(T)}{Var(T) + Var(\zeta_t)} \tag{5}$$

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 not explained by the trait).

$$OS(Y^*_{it}) = 1 - Con(Y^*_{it}) = \frac{Var(\zeta_t)}{Var(T) + Var(\zeta_t)} \tag{6}$$

As state residual variances $Var(\zeta_t)$ were set equal across time, $Con(Y^*_{it})$ and $OS(Y^*_{it})$ are constant across time (as well as across item parcels i).

Latent state-trait (LST) models with varying means *[Jana: flesh out this model]*

Correlations between tasks *[Jana: Description of how the values for correlations between tasks were computed]*

Estimation

[Jana: quick check if this is still accurate]

Models were estimated with MPlus version 8.4 (L. K. Muthén & Muthén, 1998-2017), using Bayesian Markov-Chain Monte-Carlo sampling, with the Mplus default priors (see simulation studies in the appendix). Using inverse gamma priors `IG(0.001, 0.001)` for LST models did not substantially change the parameter estimates (see simulation study). Therefore, only the results using 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 was well below a threshold of 1.01 for the first

time after the minimum number of iterations was reached. Model syntax can be accessed by locating the respective model in the folder `writing/supplement/saves/` in the GitHub repository and opening the `.out` file using a text editor.

Model fit was evaluated by computing Posterior Predicted P-values (PPP). PPP is the probability that the newly generated data are more extreme than the observed data, as measured by a specific test statistic or discrepancy function, in this case the chi-square fit function (that is, the likelihood ratio test between the specified structural equation model and an unrestricted mean and variance covariance model), see Asparouhov & Muthén (2010). The PPP is computed via the following steps: For a given MCMC iteration, a new data set is generated based on the model and the parameters of that iteration. Then the likelihood ratio chi-square test is applied to the real data as well as the newly generated data set to compute a fit index. The indices for the data and the generated data are then compared in size. If the value for the data is smaller, it is scored as 1 and if not, as 0. Averaging across these scores for the different iterations yields the PPP. Thus, values around .5 suggest a good model fit (no systematic difference between real and generated data) and very high and very low values suggest a poor model fit and / or model misspecification. In addition, we report the 95% CI of the difference between predicted and observed chi-square values, which should be centered around 0 for a good model fit.

In Mplus, every 10th iteration after burn-in is used to compute the PPP and the underlying continuous response variables Y^* are used to compute the PPP in case of ordinal data.

Projection predictive inference

The goal of this analysis was to select the predictor variables that are relevant for predicting performance in the different cognitive tasks over time. The selection of relevant predictor variables constitutes a variable selection problem, for which a range of different methods are available (e.g., shrinkage priors). We chose to use projection predictive inference because it is a state-of-the-art variable selection procedure that provides an excellent trade-off between model complexity and accuracy (Piironen & Vehtari, 2017), especially when the goal is to identify a minimal subset of predictors that yield a good predictive model (Pavone, Piironen, Bürkner, & Vehtari, 2020).

An overview of different projection techniques and an introduction to the projection prediction approach for generalized linear models can be found in Piironen et al. (2020). In this work, we use the extension to the generalized linear multilevel model case provided by Catalina, Bürkner, & Vehtari (2020).
The projection prediction approach can be viewed as a two-step process: The first step consists of building the best predictive model possible, called the reference model. In the context of this work, the reference model is a Bayesian multilevel regression model (repeated measurements nested in apes), including all available predictors. The reference model serves as a performance goal regarding the predictive quality for the smaller models constructed by the projection prediction procedure.

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 step-wise addition of predictors that decrease the Kullback-Leibler (KL) divergence from the reference model to the projected model. Let the reference model and the projected model have parameters θ' and θ respectively. Then by the definition of the KL divergence, the following optimization problem is obtained:

$$\begin{aligned} \theta_\perp &= \arg\min_\theta \frac{1}{n} \sum_{i=1}^{n} \mathrm{KL}\left(p\left(\tilde{y}_i \mid \theta'\right) \| p\left(\tilde{y}_i \mid \theta\right)\right) \\ &= \arg\min_\theta \frac{1}{n} \sum_{i=1}^{n} p\left(\tilde{y}_i \mid \theta'\right) \cdot \log\left(\frac{p\left(\tilde{y}_i \mid \theta'\right)}{p\left(\tilde{y}_i \mid \theta\right)}\right) \end{aligned} \tag{7}$$

The result of the projection is a list containing the best model for each number of predictors from which the final model is selected by inspecting the mean log-predictive density (`elpd`) and root-mean-squared error (`rmse`). The projected model with the smallest number of predictors which shows similar predictive performance as the reference model is chosen.

We built separate (Bayesian binomial multilevel regression) reference models for each task and ran them through the above-described projection prediction approach. The dependent variable for each task was the number of correctly solved trials in relation to the number of trials (i.e. the probability of solving a trial) per time point and task (`brms` notation: `p | trials(n) ~ predictors`, with `p` being the number of correct trials and `n` the number of trials).

The self-ordered-search task had a structurally different dependent variable, namely a count variable for each trial (0,1 or 2 redundant searches). Modeling this data would have required a poisson model. Unfortunately, we were unable to implement such a model in the projection prediction approach. Thus, we transformed the dependent variable for the self-ordered-search task so that no redundant search was coded as "correct" and one or more redundant searches were coded as incorrect. This allowed us to analyse this task in the same way as all the others and makes the results comparable. However, we are aware that we loose information through this transformation.

Continuous predictors were centered when needed. We transformed the `rank` variable into a relative rank, where a rank of value one depicts a subject with the highest possible rank. All models also included `time_point` as a predictor to assess changes that are related to time and thus task experience (learning or habituation). All reference models converged well, having no divergent transitions, R-hat values equal to 1, and large ESSs for virtually all parameters. The R-hat value is a diagnostic value to investigate the convergence of the model and refers to the same concept as the potential scale reduction (PSR) factor defined above. R-hat values close to 1 indicate that the chains have mixed well (the estimates of the chains agree with each other), while values above 1 indicate that the chains did not converge to the same value. For chains of autocorrelated samples, the effective sample size (ESS) is an estimate for the number of independent samples within a chain containing the same amount of information about the dependent variable.

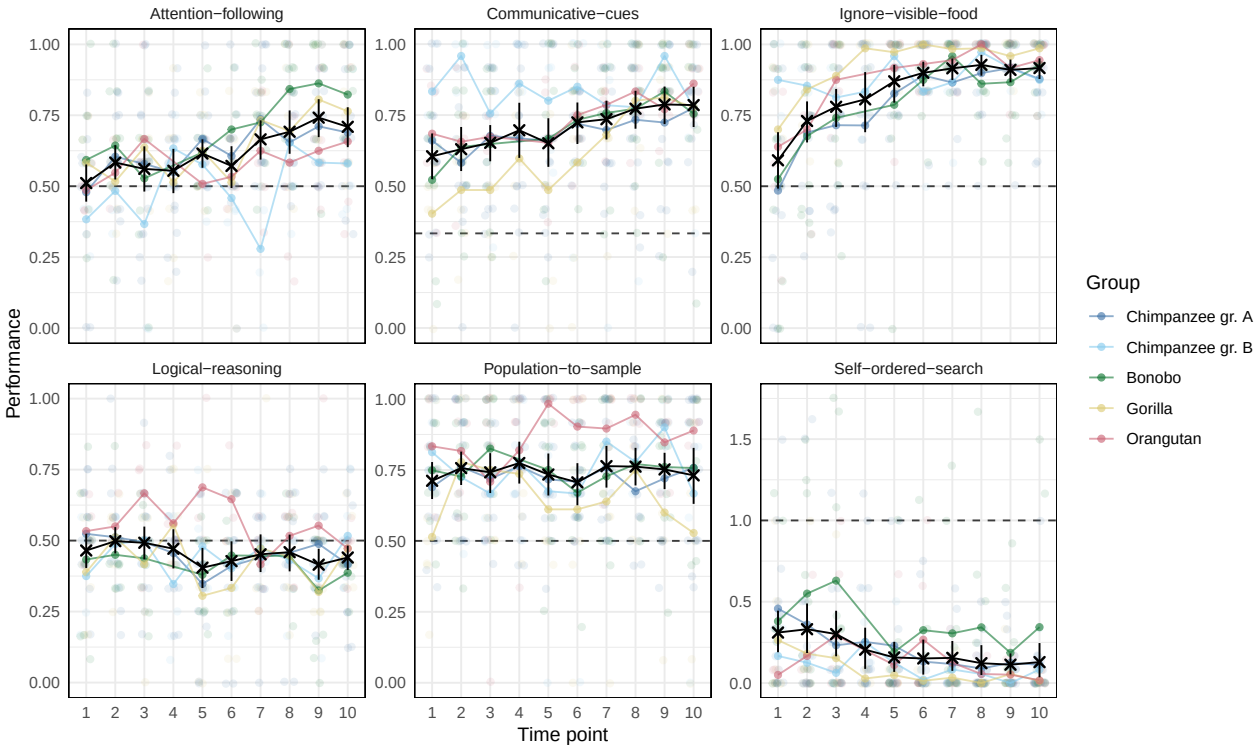
Following step two, we performed projection prediction for each reference model separately, thus resulting in different rankings of the relevant predictors for each task. We used the R package `projpred` (Piironen, Paasiniemi, Catalina, Weber, & Vehtari, 2022), which implements the aforementioned projection prediction technique. The predictor relevance ranking is measured by the LOO cross-validated mean log-predictive density and root-mean-squared error. To find the optimal submodel size, we inspected – in line with the authors' recommendations – summaries and the plotted trajectories of the calculated `elpd` and `rmse`.

The order of relevance for the predictors and the random intercept (together called terms) is created by performing forward search. The term that decreases the KL divergence between the reference model's predictions and the projection's predictions the most goes into the ranking first. Forward search is then repeated N times to get a more robust selection. We chose the final model by inspecting the predictive utility of each projection. To be precise, we chose the model with p terms where p depicts the number of terms at the cutoff between the term that increases the `elpd` and the term that does not increase the `elpd` by any significant amount. In order to get a useful predictor ranking, we manually delayed the random intercept term to the last position in the predictor selection process. The random intercept delay is needed because if the random intercept were not delayed, it would soak up almost all of the variance of the dependent variable before the predictors are allowed to explain some amount of the variance themselves. One could have used the function `suggest_size` as a heuristic decision rule to find the optimal submodel as an alternative to a graphical inspection. However, this is not yet possible due to the delay of the random intercept term.

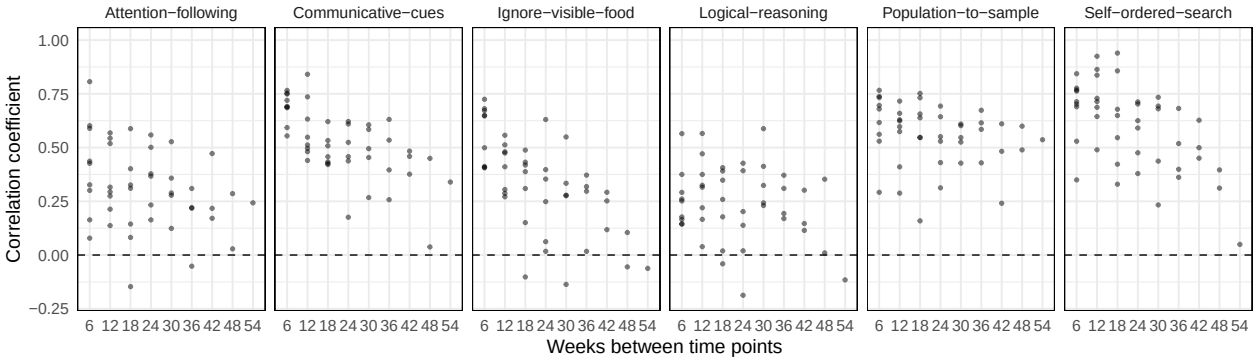
Results

Robustness, Stability and Reliability

To get an overview of the results, we first visualized the data. Supplementary Figure @ref(fig:perfplot) shows performance at the different time points. From a task-level perspective, we can say that performance was consistently above chance in the communicative-cues, ignore-visible-food and population-to-sample tasks. For attention-following, this was the case only from time point 7 onward and for logical-reasoning, performance was, if anything, below chance. For the self-ordered-search task, performance was below chance but here lower values reflect better performance (i.e. systematic avoidance of the visible food item). For attention-following, ignore-visible-food, communicative-cues and self-ordered-search there was a steady improvement in



Supplementary Figure 9: Results from the six cognitive tasks across time points. Black crosses show mean performance at each time point across species (with 95% CI). The sample size varied between time points and can be found in Supplementary Figure 1. Colored dots show mean performance by species. Dashed line shows the chance level whenever applicable.



Supplementary Figure 10: Re-test correlation coefficients plotted against the temporal distance between the testing time points.

performance over time.

For a first glimpse on the stability of individual differences, we correlated performance at the different time points for each task (all possible combinations of time points). Supplementary Figure @ref(fig:relplot) visualizes these re-test correlation coefficients against the temporal distance between time points. Correlations between time points were exclusively larger than zero for direct communicative-cues, population-to-sample and self-ordered-search. For attention-following, ignore-visible-food and logical-reasoning some correlations were negative. For all tasks, correlations between time points were lower for time points that were further apart (see also Uher, 2011). However, these re-test correlations confound measurement precision (reliability) and stability of individual differences. That is, low correlations could reflect high measurement error or a lack of stability of individual differences. We tease these components apart in the SEM models reported below.

Next, we report the SEM results for the different tasks and the relations between them. All models showed acceptable fit indices (see Supplementary Table @ref(tab:semt)). The threshold parameters for each model are shown in Supplementary Table @ref(tab:thresht).

Supplementary Table 1: Model fit indices

Task	Model	PPP	Chi 95% CI
Attention-following	LSM	0.211	-41.93 ; 106.47
	LSTM	0.301	-52.12 ; 90.63
	LSTM-V	0.083	-22.11 ; 128.34
Communicative-cues	LSM	0.596	-77.30 ; 56.69
	LSTM	0.210	-40.98 ; 98.67
	LSTM-V	0.164	-34.42 ; 107.38
Population-to-sample	LSM	0.135	-33.34 ; 118.05
	LSTM	0.495	-71.33 ; 71.92
	LSTM-V	0.482	-65.89 ; 72.54
Logical-reasoning	LSM	0.350	-55.19 ; 87.24
	LSTM	0.346	-54.41 ; 88.04
	LSTM-V	0.372	-56.06 ; 80.56
Self-ordered-search	LSM	0.451	-63.99 ; 81.44
	LSTM	0.322	-54.66 ; 91.44
	LSTM-V	0.279	-48.95 ; 95.72
Ignore-visible-food	LSM	0.498	-71.69 ; 72.83
	LSTM	0.313	-55.50 ; 91.36
	LSTM-V	0.256	-47.53 ; 97.14

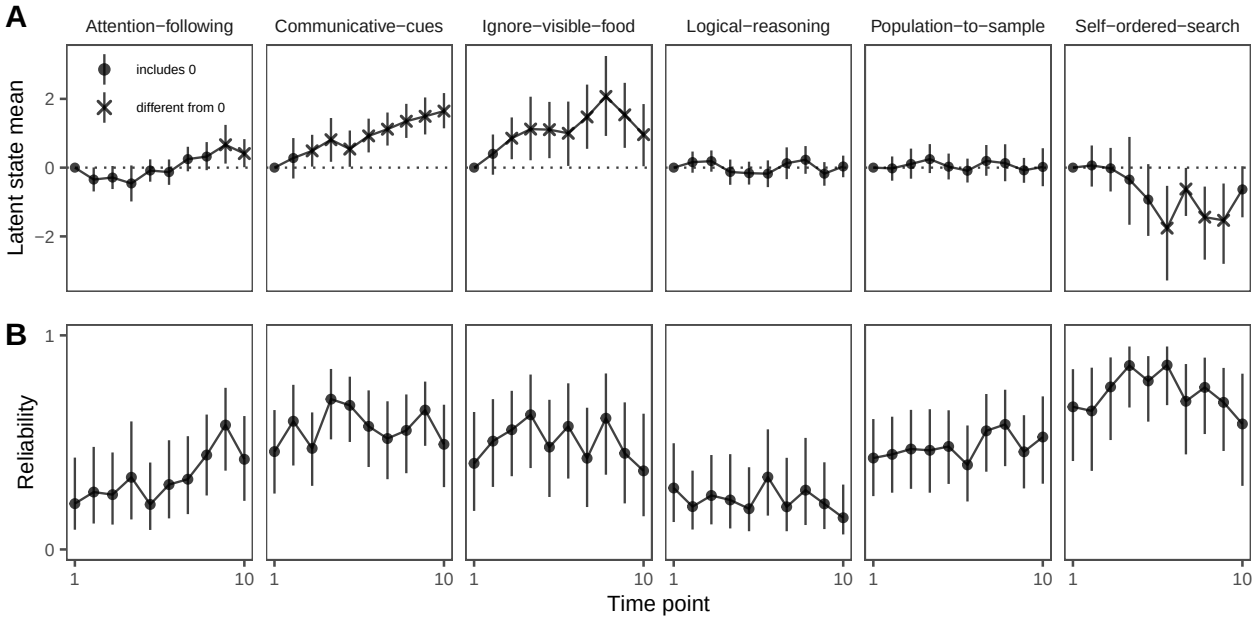
Note:
LSM = Latent state model
LSTM = Latent state-trait model
LSTM-V = Latent state-trait model with varying means
PPP = Posterior predictive p-value. Cut-off values (two-tailed): 0.025 and 0.975
Chi 95% CI = 95%CI of difference between predicted and observed chi-square values

Supplementary Table 2: Threshold parameters

Task	Model	T1	T2	T3	T4	T5
Attention-following	LSM	-1.109	-0.204	0.404	1.074	
	LSTM	-1.202	0.004	0.975	1.972	
	LSTM-V	-1.109	-0.204	0.404	1.074	
Communicative-cues	LSM	-0.946	-0.156	0.854	1.946	

	LSTM	-2.08	-1.189	-0.097	1.056	
	LSTM-V	-1.885	-1.065	-0.035	1.062	
Population-to-sample	LSM	-1.581	-0.629	-0.191	0.318	
	LSTM	-2.642	-1.22	-0.2	1.043	
	LSTM-V	-2.487	-1.141	-0.186	0.97	
Logical-reasoning	LSM	-0.975	-0.259	0.606	1.333	
	LSTM	-0.975	-0.259	0.606	1.333	
	LSTM-V	-0.975	-0.259	0.606	1.333	
Self-ordered-search	LSM	0.424	0.765	1.178	1.441	
	LSTM	0.533	0.966	1.394		
	LSTM-V	0.533	0.966	1.394		
Ignore-visible-food	LSM	-1.306	-0.83	-0.341	0.25	1.117
	LSTM	-1.946	-1.408	-0.874	-0.231	0.698
	LSTM-V	-1.584	-1.122	-0.651	-0.085	0.742

Note:
LSM = Latent state model
LSTM = Latent state-trait model
LSTM-V = Latent state-trait model with varying means
T1-5 = Threshold parameters for response categories

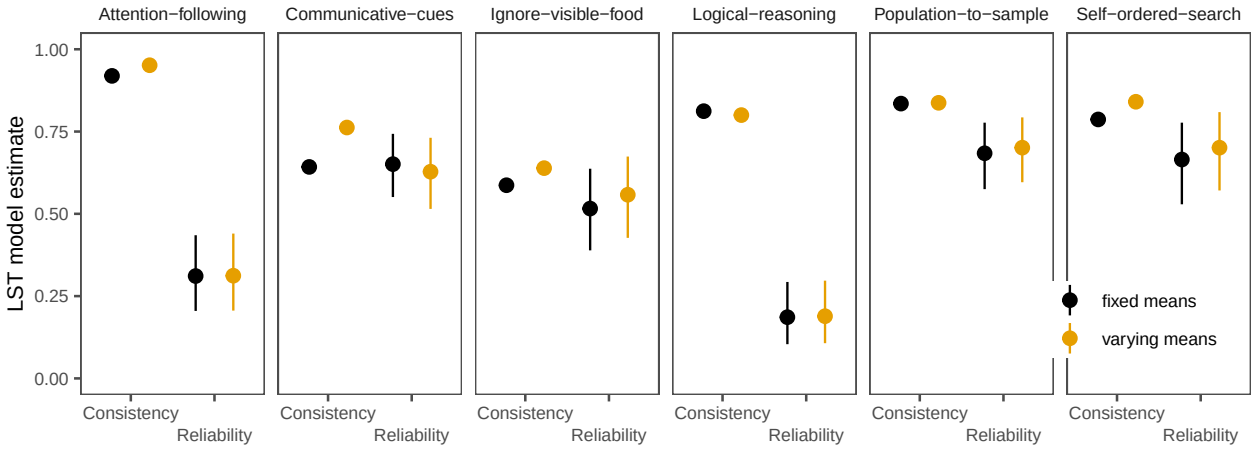


Supplementary Figure 11: Latent means and reliability estimates with 95% CI for each time point based on LSM. The sample size varied between time points and can be found in Supplementary Figure 1. Means at time point 1 are set to 0.

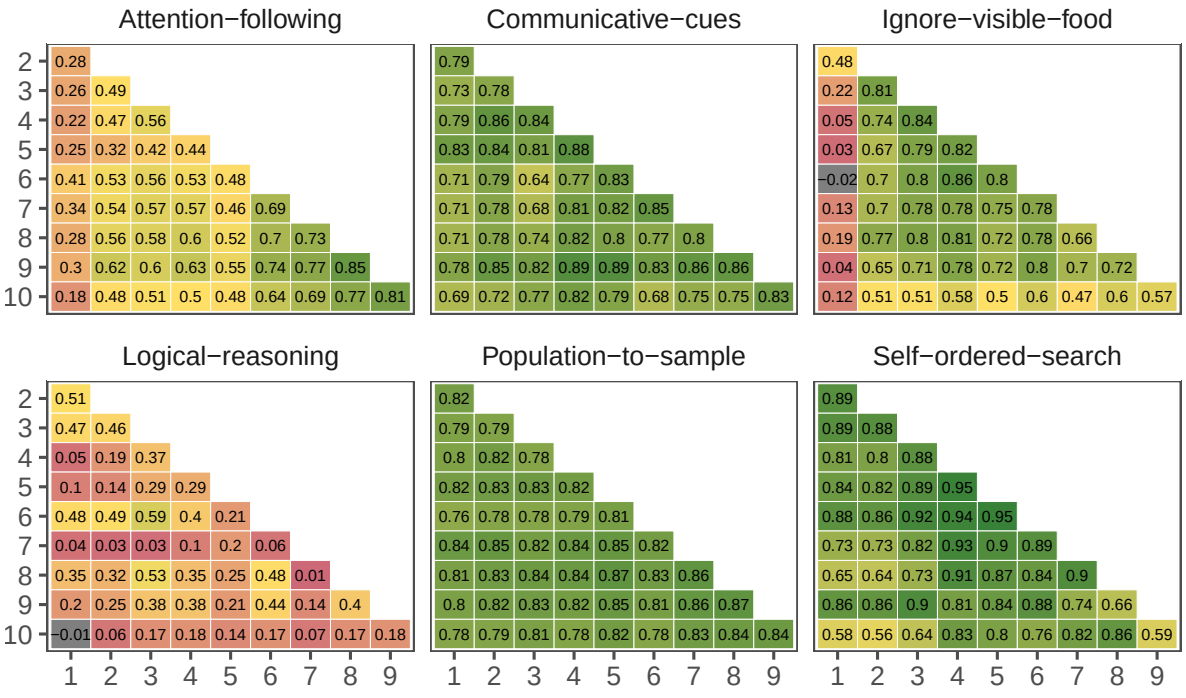
[Jana: The way we compute consistency does not give a confidence interval for the LST model with fixed or varying means. Any chance we can get this? In the last paper we had this at least for the model with the fixed means]

Attention-following

To fit the models, the response categories of 0 or 1 solved trials had to be collapsed into one category due to sparsity. Furthermore, the thresholds could not be set equal for test-half 2 at time point 1 and 9 as well as



Supplementary Figure 12: Posterior mean for model parameters (with 95% CrI) from LSTM for the four tasks based on data.



Supplementary Figure 13: Correlations between latent state variables based on LSM for the different tasks.

test-half 1 at time point 4 due to a different number of observed categories for the respective test halves and time point combination. Latent means can still be compared across time for the state factors based on the respective other test half.

Supplementary Figure @ref(fig:lseplot) visualizes the latent state means and reliability estimates from the LS model. Reliability was low for earlier time points and increased towards the end of the study. Latent means also slightly increased throughout the study and were significantly different from zero (i.e. time point 1) at time point 9 and 10. Supplementary Figure @ref(fig:lsplot) gives the correlations between the latent states for the different time points. Correlations between latent states were moderate to low for time points below 7 and substantially higher from then on. Taken together, this pattern suggests substantial re-organization of individual differences that stabilized towards the end of the study.

In the LST model with fixed means, the consistency coefficient was estimated to be 0.92 which was very similar to the model with variable means (0.95). This means that more than 90% of true inter-individual differences are attributable to stable (trait) differences between individuals (**Jana?** das beißt sich mit dem LS model - hast Du eine Idee warum das so ist?). Reliability (across time points) was estimated to be low (fixed means: 0.31; variable means: fixed means: 0.31) (see Supplementary Figure @ref(fig:lsteplot)). However, when only considering the data from time point 7 and later, reliability was substantially higher (0.66)

Taken together, there was considerable variation both in group- and individual-level performance – particularly in the earlier time points. Towards the end of the study, group-level performance stabilized on a level above chance (see Supplementary Figure @ref(fig:perfplot)) and individual-differences also became more stable. Presumably, some individuals changed the way they approached the task half-way through the study – seeing it initially as a competitive and only later as a cooperative task – so that systematic individual differences only emerged at later time points.

Communicative-cues

The lowest two categories (0 and 1 correctly solved trials per test half) were collapsed due to sparsity. Thresholds could not be set equal for test half 2 at time point 2, due to a different number of observed categories for the respective test half and time-point combination. Latent means can still be compared across time.

The latent state means in the LS model steadily increased over the course of the study (see Supplementary Figure @ref(fig:lseplot)). Reliability fluctuated around a moderate level (range: 0.46 – 0.7). The correlations between latent states for the different time points were nevertheless generally high (Supplementary Figure @ref(fig:lsplot)).

In the LST model with fixed means, the consistency coefficient was estimated to be 0.64 which was somewhat lower compared to the model with variable means (0.76). This means that ~ 60% to 70% of true inter-individual differences are attributable to stable (trait) differences between individuals. Reliability (across time points) was estimated to be moderate to acceptable (fixed means: 0.65; variable means: fixed means: 0.63) (see Supplementary Figure @ref(fig:lsteplot)).

Taken together, while there was a steady increase in group-level performance (see Supplementary Figure @ref(fig:perfplot)), individual-differences remained fairly stable. This suggests an overall learning effect that was more or less the same across individuals.

Ignore-visible-food

The lowest two categories (0 and 1 correctly solved trials per test half) were collapsed due to sparsity. Thresholds could not be set equal for test half 1 at time point 7 and for test half 2 at time point 8 and 10, due to a different number of observed categories for the respective test half and time-point combination. Latent means can still be compared across time.

In the LS model, the latent state means steadily increased until time point 8 and decreased again for the last two time points. Yet, the latent mean at time point 10 was still significantly different from 0 (i.e. time point 1, see Supplementary Figure @ref(fig:lseplot)). Note, however, that the absolute level of performance

was close to ceiling from time point 6 onward. Reliability fluctuated around a moderate level (range: 0.37 – 0.63). The correlations between latent states for the different time points were high (Supplementary Figure @ref(fig:lsplot)), with the exception of time point 1, which did not correlate with most of the other time points.

In the LST model with fixed means, the consistency coefficient was estimated to be 0.59 which was very similar to the model with variable means (0.64). This means that around 60% of true inter-individual differences are attributable to stable (trait) differences between individuals. Reliability (across time points) was estimated to be moderate (fixed means: 0.52; variable means: 0.56) (see Supplementary Figure @ref(fig:lsteplot)).

Taken together, there was a steady increase in group-level performance (see Supplementary Figure @ref(fig:perfplot)) over time. Except of time point 1, individual-differences were fairly stable, suggesting a a fairly consistent learning effect across individuals from time point 2 onward.

Logical-reasoning

The lowest two (0 and 1 correctly solved trials per test half) as well as the highest two categories (5 and 6 solved trials) were collapsed due to sparsity. Thresholds could not be set equal for test half 1 at time points 7, 8 and 9, due to a different number of observed categories for the respective test half and time-point combination. Latent means can still be compared across time.

There was no change in latent state means (see Supplementary Figure @ref(fig:lseplot)). Reliability fluctuated on a low level (range: 0.15 – 0.34). The correlations between latent states for the different time points were generally low, some even negative (Supplementary Figure @ref(fig:lsplot)).

In the LST model with fixed means, the consistency coefficient was estimated to be 0.81 which was very similar to the model with variable means (0.8). This means that around 80% of true inter-individual differences are attributable to stable (trait) differences between individuals. However, the reliability (across time points) was estimated to be very low, making the consistency estimate difficult to interpret (fixed means: 0.19; variable means: 0.19) (see Supplementary Figure @ref(fig:lsteplot)).

In combination with the finding that group-level performance was not reliably different from chance (see Supplementary Figure @ref(fig:perfplot)), the low correlations between latent states and low reliability estimates suggest that this task did not measure any meaningful individual-differences in logical reasoning abilities. For this reason, we did not include this task when predicting cognitive performance by external predictors.

Population-to-sample

The lowest two categories (0 and 1 correctly solved trials per test half) were collapsed due to sparsity. Thresholds could not be set equal for test half 1 at time point 3 and for test half 2 at time point 8 and 10, due to a different number of observed categories for the respective test half and time-point combination. Latent means can still be compared across time.

In the LS model, the latent state means did not change over time (see Supplementary Figure @ref(fig:lseplot)). Reliability varied on a low to moderate level (range: 0.4 – 0.58). The correlations between latent states for the different time points were generally high (Supplementary Figure @ref(fig:lsplot)).

In the LST model with fixed means, the consistency coefficient was estimated to be 0.83 which was very similar to the model with variable means (0.84). This means that around 80% of true inter-individual differences are attributable to stable (trait) differences between individuals. Reliability (across time points) was estimated to be acceptable (fixed means: 0.68; variable means: fixed means: 0.7) (see Supplementary Figure @ref(fig:lsteplot)).

In sum, the results suggest that task-level results were robust and individual-level performance was stable over time. As noted above – and as can be seen in Supplementary Figure @ref(fig:perfplot) – performance on a task level was clearly above chance in all sessions.

Self-ordered-search

This task had a different response pattern compared to the other tasks. The maximum score for each trial was 2 and not 1. Furthermore, higher scores indicate worse performance (i.e. more errors). All scores of 4 and higher were collapsed into a single category due to sparsity. Thresholds could not be set equal for test half 1 at time point 4 and for test half 2 at time point 4, 5 and 10, due to a different number of observed categories for the respective test half and time-point combination. Latent means can still be compared across time except for time point 4.

From time point 6 onwards, the latent means were significanlty lower compared to the first time point (i.e. subjects made fewer errors, see Supplementary Figure @ref(fig:lseplot)). Reliability varied between moderate and good (range: 0.59 – 0.86). The correlations between latent states for the different time points were generally high (Supplementary Figure @ref(fig:lsplot)).

In the LST model with fixed means, the consistency coefficient was estimated to be 0.79 which was very similar to the model with variable means (0.84). This means that around 80% of true inter-individual differences are attributable to stable (trait) differences between individuals. Reliability (across time points) was estimated to be acceptable (fixed means: 0.66; variable means: 0.7) (see Supplementary Figure @ref(fig:lsteplot)).

Taken together, the results, suggest stable individual-differences with a slight group-level decrease in error rate (see Supplementary Figure@ref(fig:perfplot)).

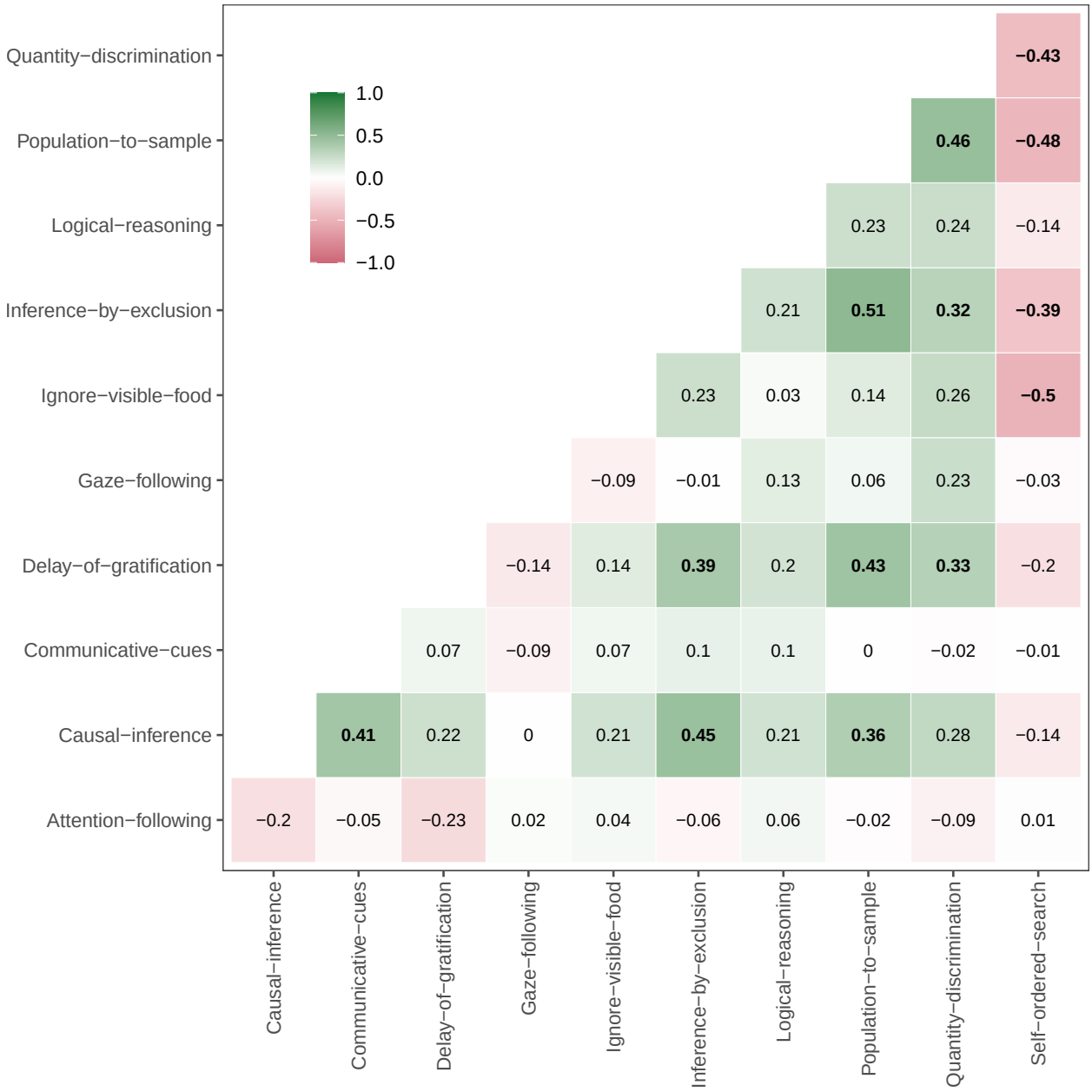
Summary

The six tasks differed substantially in what they revealed about group- and individual-level variation. What stands out is the widespread change in performance over time. For all tasks except population-to-sample and logical-reasoning we observed an improvement in performance over time. This group-level change, however, has different individual-level interpretations for the different tasks. For communicative-cues, ignore-visible-food and self-ordered-search, individual differences remained relatively stable despite the group-level change suggesting stable individual differences combined with a systematic learning effect across individuals. In contrast, for attention-following, there was little stability in individual differences at earlier time points and only towards the end emerged a more stable ordering of individuals. In combination with the low reliability at earlier time points, this suggests that at least some individuals changed their response strategy in the course of the study. The combination of low reliability, chance-level performance and low correlation of latent states for logical-reasoning suggests that this task is not suited to probe individual differences in logical reasoning abilities in great apes. It is also noteworthy that the reliability estimates are on average lower compared to a previous study testing the same individuals on different tasks (Bohn et al., 2023). One explanation might be the increase in performance over time. At the beginning of the study, more individuals might have chosen randomly instead of using the available information provided in the task setup and the demonstrations. By definition, random variation is not reliable. With time, more and more individuals started using the available information so that inter-individual differences in how good they are in using it could be detected.

Relations between tasks

Supplementary Figure@ref(fig:mtmplot) shows the correlations between trait estimates for the different tasks reported in the present study as well as those reported in Bohn et al. (2023). For the tasks reported in Bohn et al. (2023) we used the data from phase 2 because it was closer in time. Overall, most correlations were not significantly different from zero (i.e. the 95% CI did include zero). Because of this low average level of correlations, we decided not to explore models with higher-order factors and will only interpret the qualitative patterns.

Conceptually, the tasks can be clustered in the following broader domains: *social cognition* (attention-following, gaze-following, communicative-cues), *reasoning about quantities* (quantity-discrimination, population-to-sample), *executive functions* (delay-of-gratification, self-ordered-search, ignore-visible-food) and *inferential reasoning* (logical-reasoning, causal-inference, inference-by-exclusion). As a first step, we will evaluate whether we find evidence for such a clustering in the data.



Supplementary Figure 14: Correlations between trait estimates. Bold correlations are different from zero as judged by the 95% CI.

There was no significant correlation between any of the social cognition tasks. Furthermore, attention-following and gaze-following did not correlate significantly with any of the other tasks and communicative-cues correlated only with causal-inference – a result we will discuss below. Thus, and in line with previous work (Herrmann, Hernández-Lloreda, Call, Hare, & Tomasello, 2010), we found no evidence for shared cognitive processes in tasks measuring different aspects of social cognition.

The two tasks measuring reasoning about quantities did correlate significantly. Both tasks require discriminating between different quantities, directly in the case of quantity-discrimination and as part of the decision making process in the case of population-to-sample. Deciding between the samples from the two populations requires discriminating between the relative quantities within each bucket from which the samples were drawn.

Within the executive functions measures, self-ordered-search and inhibit-visible-food were significantly correlated but none of the two correlated with delay-of-gratification. The significant correlation can be explained by the need to inhibit a premature response (selecting visible food or a cup that was previously rewarded) in both tasks. It has been argued that delay-of-gratification requires self-control (tolerating a longer waiting time to gain a more valuable reward) over and above behavioral inhibition (Beran, 2015). From this point of view, individual differences in the delay-of-gratification task might be due to differences in self control and less due to differences in inhibition.

Finally, for the three inferential reasoning measures we found a correlation between inference-by-exclusion and causal-inference. Logical-reasoning did not correlate with either (neither did it with any other task). This is not surprising given the results reported above: the observed variation in the logical-reasoning task was largely noise and did not reflect systematic individual differences. The correlation between causal-inference and inference-by-exclusion is most likely due to the fact that both tasks involve making inferences about the location of food based on reasoning about its physical properties.

Next we turn to the correlations across domains. Perhaps the most surprising finding is the correlation between causal-inference and communicative-cues. On a closer look, the origin might be task impurity in that there are two ways to solve the causal-inference task: first, as hypothesized, by using the rattling sound to infer the location of the food. Second, by interpreting the experimenter's shaking of the cup as a communicative cue, which is very similar to the communicative-cues task. Thus, we suspect that at a least some individuals solved the task via the second route.

Finally, there was a cluster of significant correlations between delay-of-gratification, self-ordered-search, inference-by-exclusion, causal-inference, population-to-sample and quantity discrimination. Of the 15 possible correlations, only four were non-significant. One commonality between these tasks that might – in part – explain this pattern is that they all benefit from sustained attention to the task. Sustained attention facilitates the processing of the experimenter's demonstrations (population-to-sample, inference-by-exclusion, causal-inference, delay-of-gratification), one's actions on the setup (self-ordered-search) or visually complex stimuli (quantity discrimination). Tentative support for this idea comes from the analysis of relevant predictors (see Bohn et al., 2023 and below) in which `time spent in research` was selected as a relevant predictor of performance for all of these tasks except causal-inference. This predictor reflects individual's experience with experimental studies, which often involve sustained attention to distributions of food items, actions of conspecifics and/or demonstrations by experimenters.

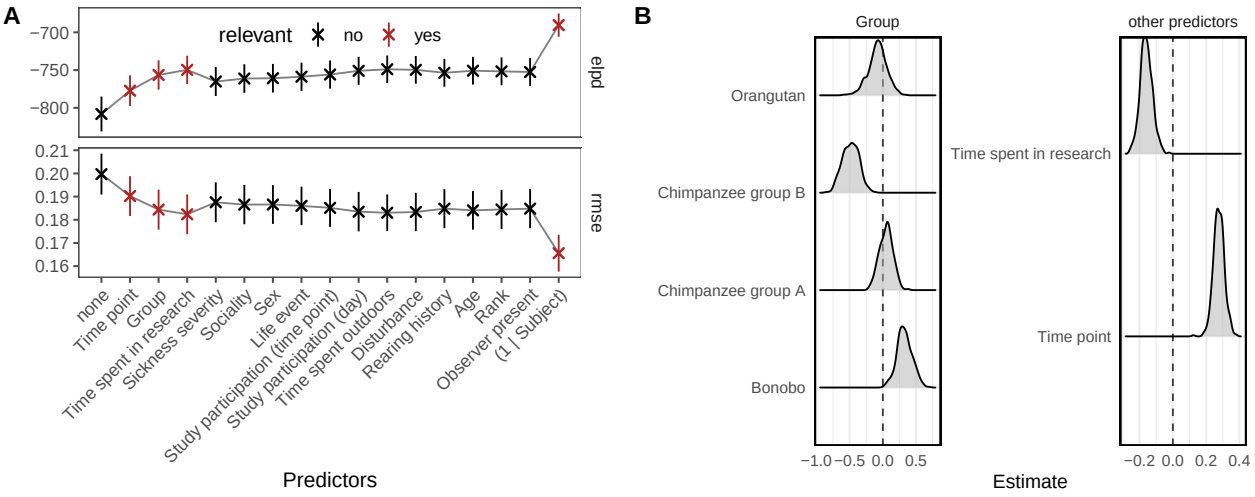
Predictability

In the following, we describe the Projection Prediction Inference results for each task. For each task, we differentiate between relevant and irrelevant predictors and report the projected posterior distribution for relevant predictors.

Attention-following

Supplementary Figure @ref(fig:attselp) visualizes the results. Out of the 13 predictor variables we analysed, we selected `time point`, `group`, and `time spent in research` to be relevant in addition to the random intercept term. When inspecting the projected posterior distribution for `time spent in research`, we found

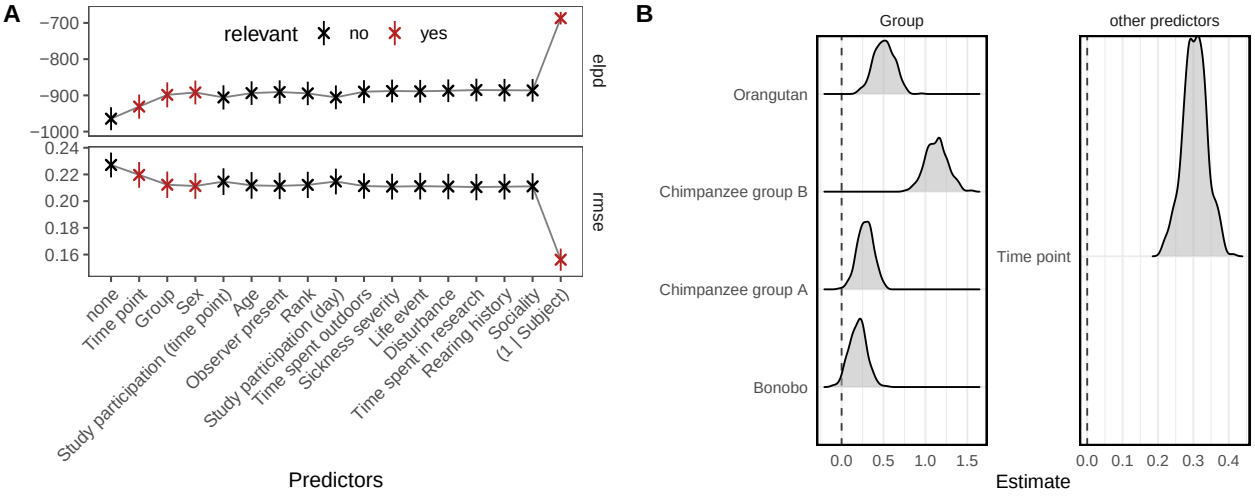
that less experienced apes performed better. With respect to `group`, in comparison to the reference group, the Gorillas, the Chimpanzee group B performed worse while the Bonobos performed better. The results for `time point` mirrored the results above in that they suggested a better performance at later time points.



Supplementary Figure 15: Predictor selection for attention-following. A) Elpd and RMSE values (with standard error) for predictors, ordered by importance (left to right) according to the cross-validated projection prediction model. Note that the random intercept term was forced to be the last one to be included. B) Projections for the selected predictors based on the submodel. Reference level for group are Gorillas.

Communicative-cues

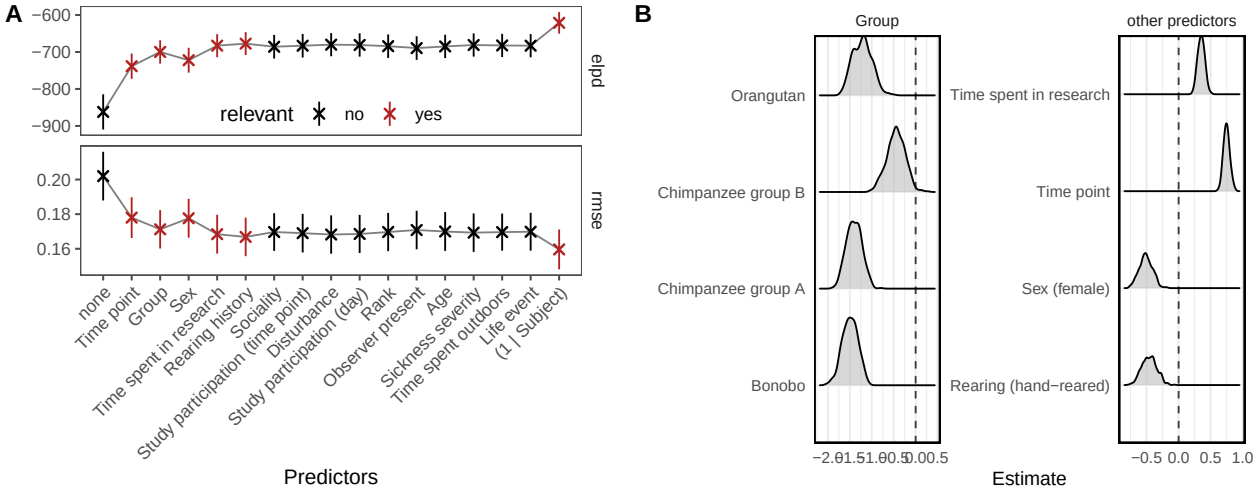
Supplementary Figure @ref(fig:comselp) visualizes the results. The predictors for the communicative-cues task were `time point` and `group`. The results for `time point` mirrored the results above in that they suggested a better performance at later time points. With respect to `group`, in comparison to the reference group, the Gorillas, all other groups performed better, in particular the Chimpanzee group B.



Supplementary Figure 16: Predictor selection for attention-following. A) Elpd and RMSE values (with standard error) for predictors, ordered by importance (left to right) according to the cross-validated projection prediction model. Note that the random intercept term was forced to be the last one to be included. B) Projections for the selected predictors based on the submodel. Reference level for group are Gorillas.

Ignore-visible-food

Supplementary Figure @ref(fig:visselp) visualizes the results. The selected predictors were `time point`, `group`, `sex`, `time spent in research` and `rearing history`. Once again, the results for `time point` mirrored the results above in that they suggested a better performance at later time points. For `group`, in comparison to the reference group, the Gorillas, all other groups performed worse. The results for `time spent in research` suggest that more experience apes performed better. Finally, female (`sex`) and hand-reared (`rearing history`) individuals performed worse.



Supplementary Figure 17: Predictor selection for ignore-visible-food. A) Elpd and RMSE values (with standard error) for predictors, ordered by importance (left to right) according to the cross-validated projection prediction model. Note that the random intercept term was forced to be the last one to be included. B) Projections for the selected predictors based on the submodel. Reference level for group are Gorillas.

Population-to-sample

Supplementary Figure @ref(fig:popselp) visualizes the results. The selected predictors were `time spent in research`, `rearing history`, `group`, `sex` and `rank`. The results for `time spent in research` suggest that more experience apes performed better. All groups except the Chimpanzee A group outperformed the Gorillas. Females (`sex`) performed better while hand-reared (`rearing history`) individuals performed worse. Finally, higher ranking (`rank`) individuals performed better.

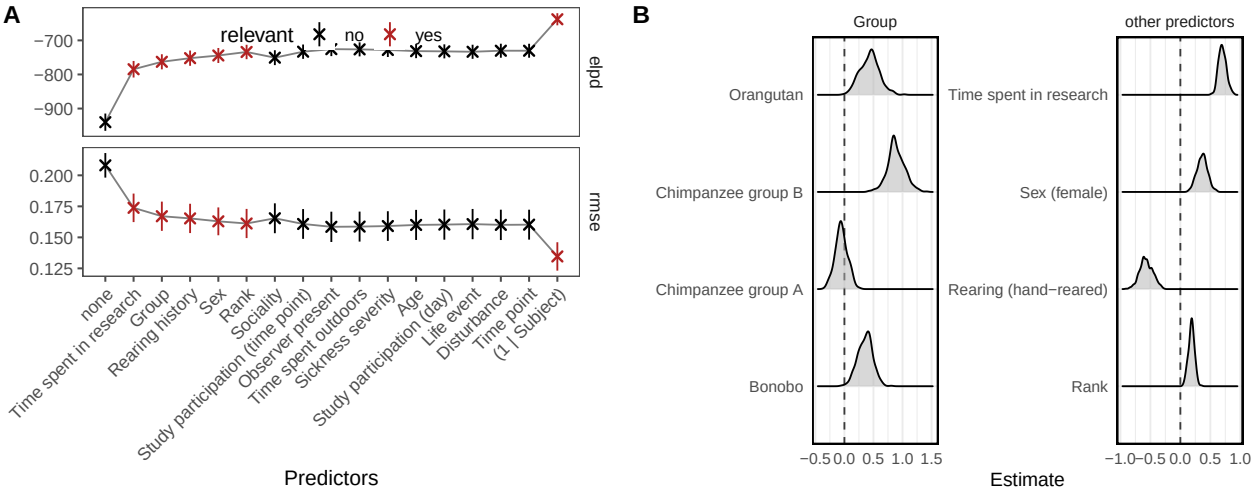
Self-ordered-search

Please note that the dependent variable for this task was transformed to match the structure of the other tasks. As described above, we coded trials with no redundant search as "correct" and trials with one or more redundant searches as "incorrect".

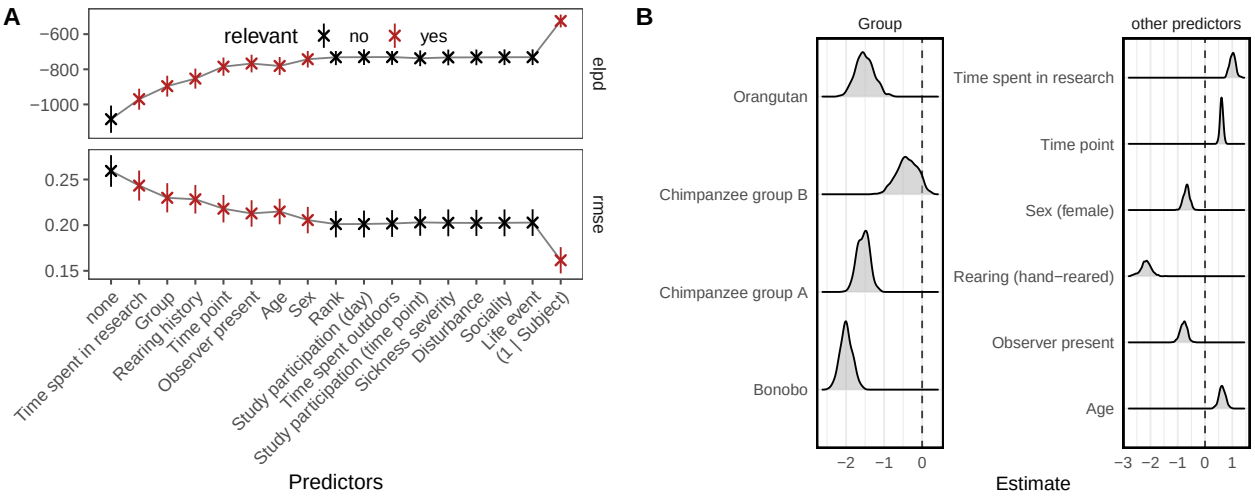
Supplementary Figure @ref(fig:inhselp) visualizes the results. The selected predictors were `time point`, `time spent in research`, `group`, `rearing history`, `Observer present`, `age`, and `sex`. Once again, performance increased over time (`time point`). More experience apes performed better (`time spent in research`). In comparison to the Gorillas, all other groups performed worse, with the Chimpanzee group B being closest. Hand-reared (`rearing history`) individuals performed worse. Performance was worse when an observer was present. Performance was better for older individuals. Finally, males (`sex`) performed better.

Summary

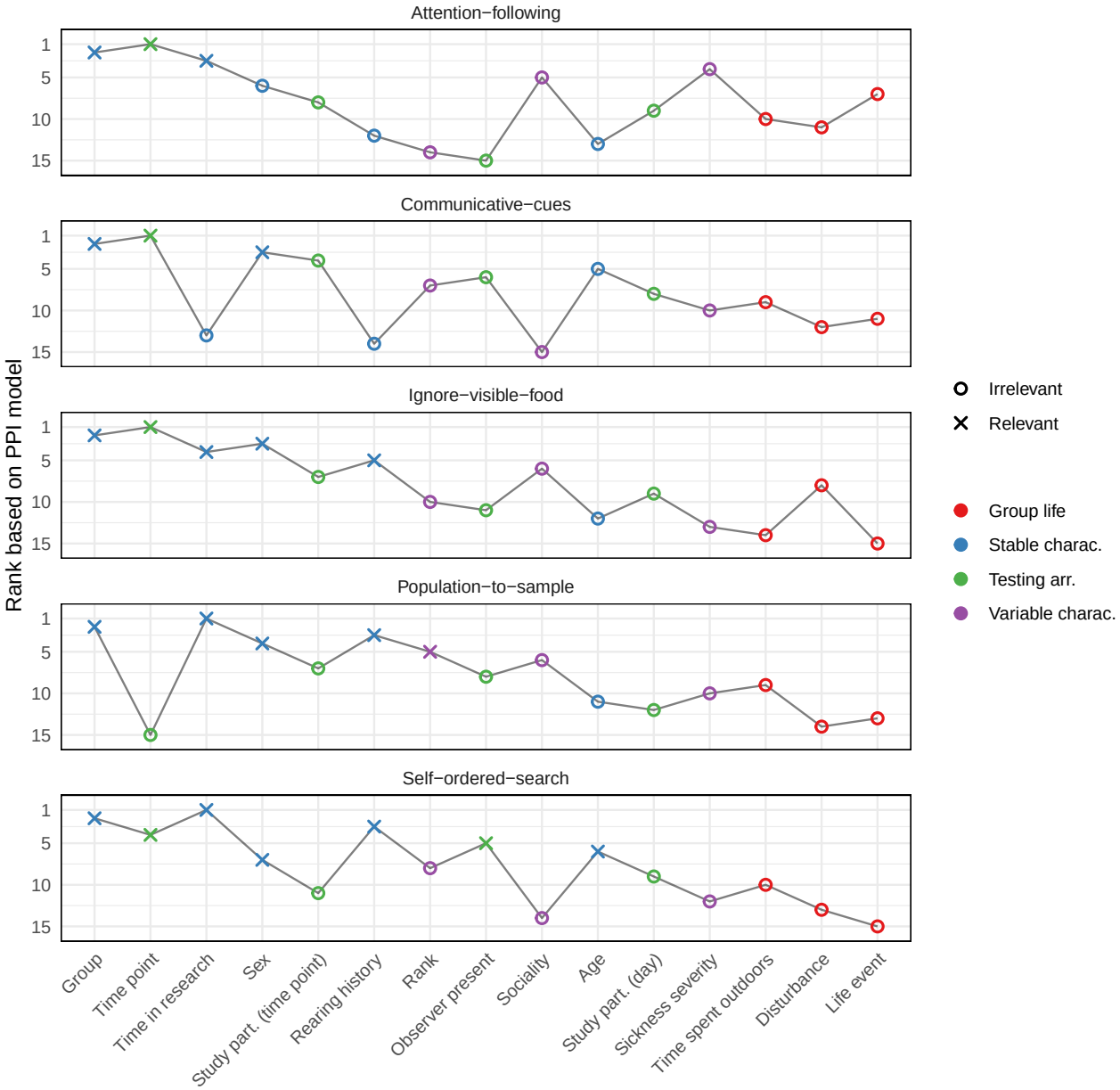
Supplementary Figure @ref(fig:ppisum) summarizes the selected predictors across tasks. Supplementary Figure @ref(fig:ppipreds) shows the projected posterior model estimates for the predictors that were selected as relevant.



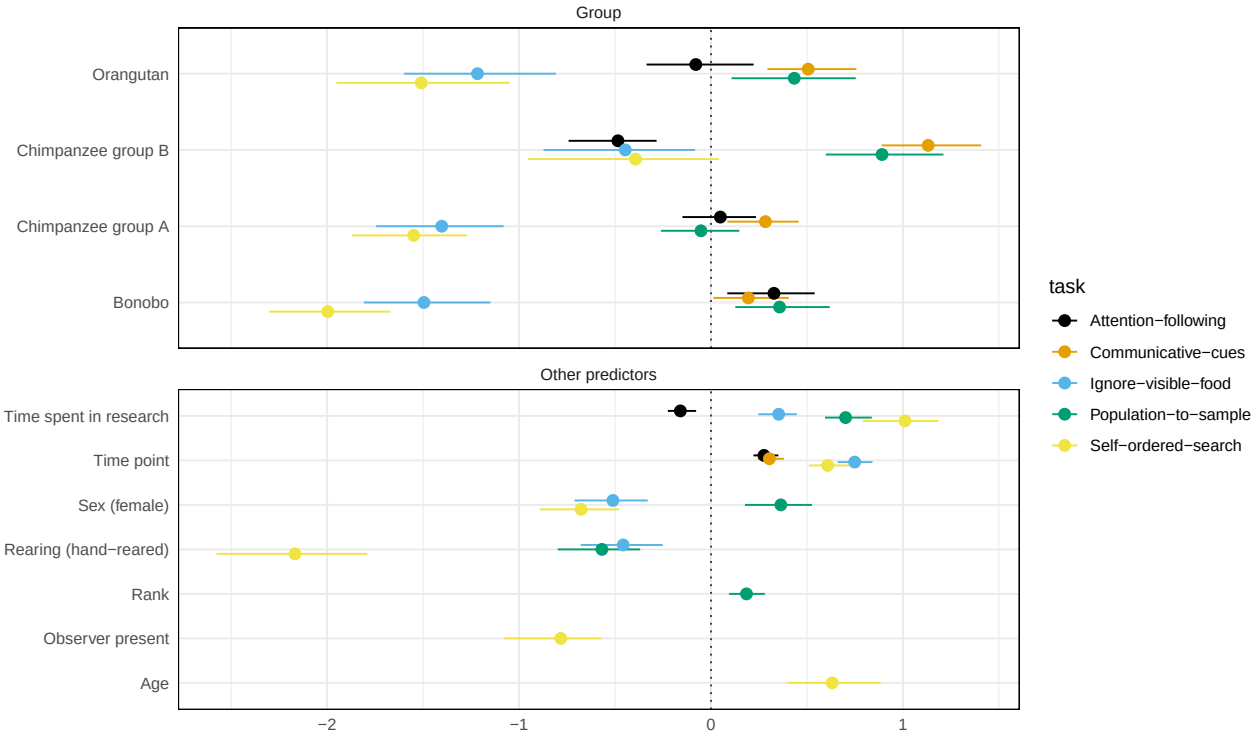
Supplementary Figure 18: Predictor selection for population-to-sample. A) Elpd and RMSE values (with standard error) for predictors, ordered by importance (left to right) according to the cross-validated projection prediction model. Note that the random intercept term was forced to be the last one to be included. B) Projections for the selected predictors based on the submodel. Reference level for group are Gorillas.



Supplementary Figure 19: Predictor selection for self-ordeered-search. A) Elpd and RMSE values (with standard error) for predictors, ordered by importance (left to right) according to the cross-validated projection prediction model. Note that the random intercept term was forced to be the last one to be included. B) Projections for the selected predictors based on the submodel. Reference level for group are Gorillas.



Supplementary Figure 20: Predictor ranking and selection based on PPI models. Crosses mark predictors that were selected to be relevant based on the PPI models. Color shows the broader category each predictor belongs to. The x-axis is sorted by the average rank across tasks.



Supplementary Figure 21: Posterior model estimates for the selected predictors for each task based on data. Points show means with 95% Credible Interval. Color denotes task. For categorical predictors, the estimate gives the difference compared to the reference level (Gorilla for group).

Across tasks, the random intercept term `(1 | subject)` was the predictor that improved model fit the most. In line with results reported by Bohn et al. (2023), this suggests that idiosyncratic developmental processes or genetic pre-dispositions, which operate on a much longer time-scale than what we captured in our study, account for a substantial portion of the variance in cognitive abilities between individuals.

However, for two tasks, other predictors had an comparable explanatory power – something we did not observe in Bohn et al. (2023). For population-to-sample, `time spent in research` improved the model fit even more than adding the random intercept at the end did. This could be interpreted that performance in this task strongly depends on having learned to pay attention to stimuli and the human experimenter. For ignore-visible-food, `time point` had an influence exceeding that of the random intercept term. We think this result reflects the strong within-task learning effect across subjects. Because performance increased substantially with time, most of the variation captured by `time point` exceeded the variation between individuals.

For the remaining predictors, the most highly-ranked and frequently selected ones came from the group of stable individual characteristics. The big exception being `time point`, which was ranked second across tasks. This pattern aligns with the SEM results, in which we saw that most of the variance in performance could be traced back to stable trait differences between individuals. The remaining occasion specific variation was largely due to continuous improvement over time, most likely reflecting task-specific learning processes. The remaining time-varying predictors did not account for much variation over and above stable trait differences and learning.

The predictor selected most often was `group`. It was the only predictor that was selected as relevant for all tasks. Differences between groups were, however, variable in that the ranking of the groups changed from task to task. For example, the Gorillas performed best in ignore-visible-food and self-ordered-search, the Chimpanzee group B performed best in communicative-cues and population-to-sample and the Bonobos performed best in attention-following. This speaks against clear species or group differences in general

cognitive performance. Again, the most likely explanation for group differences is an interaction between species-specific dispositions and individual- / task-level developmental processes.

The predictors that were selected more than once influenced performance in variable ways. As mentioned above, `time point` always had a positive effect because performance increased with time. Whenever `rearing` was selected to be relevant, mother-reared individuals outperformed others. `Time spent in research` had a positive effect, suggesting that more experience with research leads to better performance, except for attention-following. The effect of `sex` was variable in that females outperformed males in population-to-sample but males outperformed females in self-ordered-search and ignore-visible-food.

Supplementary References

Asparouhov, T., & Muthén, B. (2010). Bayesian analysis in mplus: Technical implementation. Mplus technical report, Version 3. Retrieved from http://www.statmodel.com/download/Bayes3.pdf

Beran, M. J. (2015). The comparative science of "self-control": What are we talking about? *Frontiers in Psychology*, *6*, 51.

Bohn, M., Eckert, J., Hanus, D., Lugauer, B., Holtmann, J., & Haun, D. B. (2023). Great ape cognition is structured by stable cognitive abilities and predicted by developmental conditions. *Nature Ecology & Evolution*, *7*(6), 927–938.

Bollen, K. A. (1989). *Structural equations with latent variables* (Vol. 210). John Wiley & Sons.

Catalina, A., Bürkner, P.-C., & Vehtari, A. (2020). Projection predictive inference for generalized linear and additive multilevel models. *arXiv Preprint arXiv:2010.06994*.

Diamond, A., Prevor, M. B., Callender, G., & Druin, D. P. (1997). Prefrontal cortex cognitive deficits in children treated early and continuously for PKU. *Monographs of the Society for Research in Child Development*, i–206.

Eckert, J., Call, J., Hermes, J., Herrmann, E., & Rakoczy, H. (2018). Intuitive statistical inferences in chimpanzees and humans follow weber's law. *Cognition*, *180*, 99–107.

Eid, M., Holtmann, J., Santangelo, P., & Ebner-Priemer, U. (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.

Eid, M., & Kutscher, T. (2014). Statistical models for analyzing stability and change in happiness. In K. Sheldon & R. Lucas (Eds.), *Stability of happiness: Theories and evidence on whether happiness can change* (pp. 261–297). Elsevier.

Geiser, C. (2020). *Longitudinal structural equation modeling with mplus: A latent state-trait perspective*. Guilford Publications.

Hanus, D., & Call, J. (2014). When maths trumps logic: Probabilistic judgements in chimpanzees. *Biology Letters*, *10*(12), 20140892.

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

Hoyle, R. H. (2012). *Handbook of structural equation modeling*. Guilford press.

Kajokaite, K., Whalen, A., Koster, J., & Perry, S. (2021). Fitness benefits of providing services to others: Grooming predicts survival in a neotropical primate. *bioRxiv*. http://doi.org/10.1101/2020.08.04.235788

Kaminski, J., Call, J., & Tomasello, M. (2004). Body orientation and face orientation: Two factors controlling apes' begging behavior from humans. *Animal Cognition*, *7*, 216–223.

Leckie, G. (2019). Multiple membership multilevel models. Retrieved from https://arxiv.org/abs/1907.04148

Meredith, W. (1993). Measurement equivalence, factor analysis and factorial equivalence. *Psychometrika*, *58*(4), 525–543.

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research*, *39*(3), 479–515.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*(1), 115–132.

Muthén, L. K., & Muthén, B. (1998-2017). *Mplus user's guide. Eighth edition*. Los Angeles, CA:Muthén & Muthén.

Pavone, F., Piironen, J., Bürkner, P.-C., & Vehtari, A. (2020). Using reference models in variable selection. Retrieved from https://arxiv.org/abs/2004.13118

Petrides, M. (1995). Impairments on nonspatial self-ordered and externally ordered working memory tasks after lesions of the mid-dorsal part of the lateral frontal cortex in the monkey. *Journal of Neuroscience*, *15*(1), 359–375.

Piironen, J., Paasiniemi, M., Catalina, A., Weber, F., & Vehtari, A. (2022). projpred: Projection predictive feature selection. Retrieved from https://mc-stan.org/projpred/

Piironen, J., Paasiniemi, M., & Vehtari, A. (2020). Projective inference in high-dimensional problems: Prediction and feature selection. *Electronic Journal of Statistics*, *14*(1), 2155–2197. http://doi.org/10.1214/20-EJS1711

Piironen, J., & Vehtari, A. (2017). Comparison of bayesian predictive methods for model selection. *Statistics and Computing*, *27*, 711–735. http://doi.org/10.1007/s11222-016-9649-y

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*, (34), 1–97.

Samejima, F. (1996). The graded response model. In W. van der Linden & R. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). New York: Springer.

Schmid, B., Karg, K., Perner, J., & Tomasello, M. (2017). Great apes are sensitive to prior reliability of an informant in a gaze following task. *PLoS One*, *12*(11), e0187451.

Snijders, T. A., & Kenny, D. A. (1999). The social relations model for family data: A multilevel approach. *Personal Relationships*, *6*(4), 471–486.

Steyer, R., Ferring, D., & Schmitt, M. J. (1992). States and traits in psychological assessment. *European Journal of Psychological Assessment*, *8*, 79–98.

Steyer, R., Mayer, A., Geiser, C., & Cole, D. A. (2015). A theory of states and traits—revised. *Annual Review of Clinical Psychology*, *11*, 71–98.

Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*(3), 393–408.

Uher, J. (2011). Individual behavioral phenotypes: An integrative meta-theoretical framework. Why "behavioral syndromes" are not analogs of "personality." *Developmental Psychobiology*, *53*(6), 521–548.

Völter, C. J., Mundry, R., Call, J., & Seed, A. M. (2019). Chimpanzees flexibly update working memory contents and show susceptibility to distraction in the self-ordered search task. *Proceedings of the Royal Society B*, *286*(1907), 20190715.

Völter, C. J., Tinklenberg, B., Call, J., & Seed, A. M. (2022). Inhibitory control and cue relevance modulate chimpanzees'(pan troglodytes) performance in a spatial foraging task. *Journal of Comparative Psychology*, *136*(2), 105.

Wark, J. D., Cronin, K. A., Niemann, T., Shender, M., Horrigan, A., Kao, A., & Ross, M. R. (2019). Monitoring the behavior and habitat use of animals to enhance welfare using the ZooMonitor app. *Animal Behavior and Cognition*, *6*, 158–167.