tbd...

Supplementary material

tbd...

Contents

Overview **1**

Methods **2**

Participants . . . 2

Setup . . . 2

Tasks . . . 2

Data collection . . . 5

Predictors . . . 5

Analytical framework **9**

Structural equation modelling . . . 9

Projection predictive inference . . . 13

Results **14**

Stability and Reliability . . . 14

Relations between tasks . . . 19

Predictability . . . 19

Summary **21**

Appendix **21**

SEM Simulations . . . 21

Overview

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

tasks - stability, reliability

predictors - predictability

Methods

Participants

A total of 43 great apes participated at least once in one of the tasks. This included 8 Bonobos (3 females, age 7.3 to 38.5), 24 Chimpanzees (18 females, age 2.6 to 55.4), 6 Gorillas (4 females, age 2.7 to 22.1), and 6 Orangutans (4 females, age 17 to 40.7). The sample size at the different time points ranged from 0 to 24. 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.

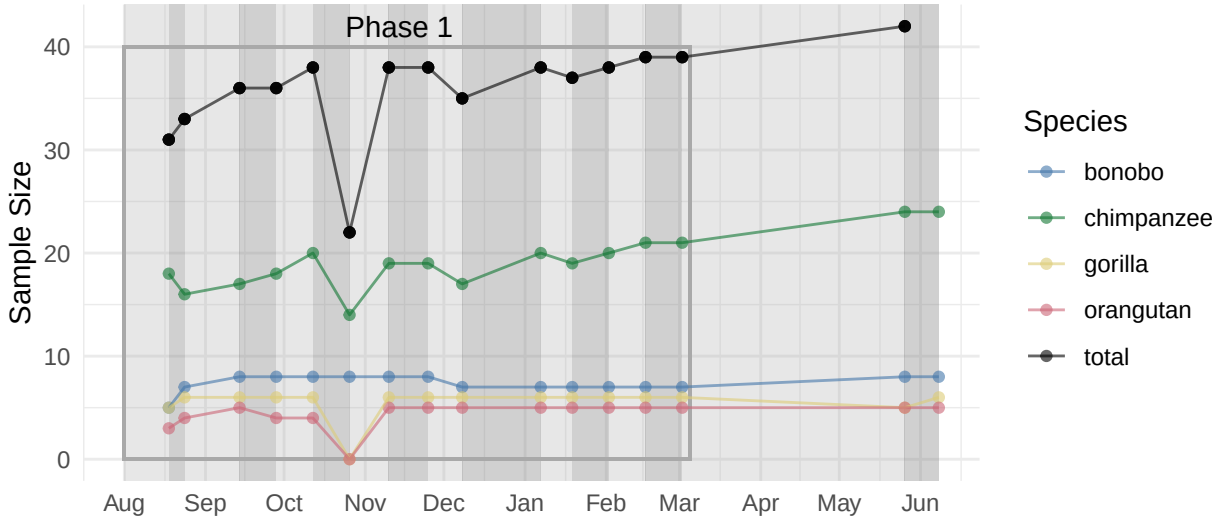

Figure S1: Sample size by species across the different time points. Time point specific predictor variables were collected during the time between two time points (shaded regions) to predict the next.

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

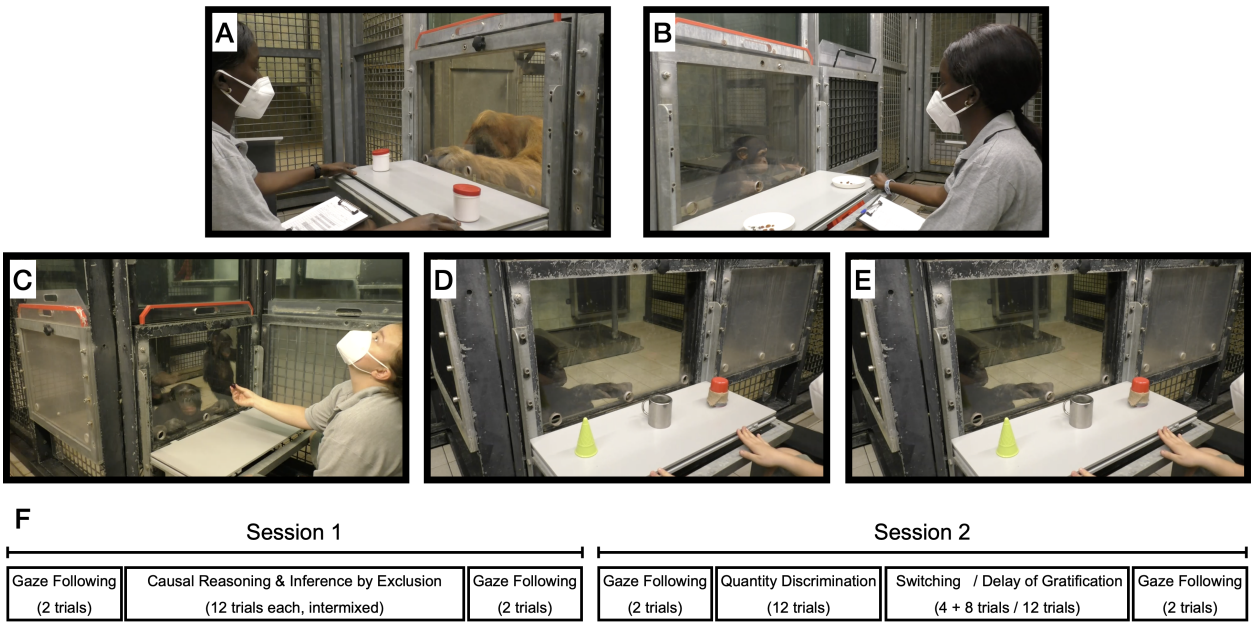

Figure S2: Setup used for the six tasks. A) Causal reasoning and inference by exclusion. B) Quantity discrimination. C) Gaze following. D) Switching. E) Delay of gratification.

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, & 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 & 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, Call, Janzen, & Levinson (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, Call, Janzen, & Levinson (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

The procedure for this task was adapted from Rosati, Stevens, Hare, & Hauser (2007). Two small plates including one and two pieces of pellet were presented left and right on the table. E moved the plate with the smaller reward forward allowing the subject to choose immediately, while the plate with the larger reward was moved forward after a delay of 20 seconds. We coded whether the subject selected the larger delayed reward (correct choice) or the smaller immediate reward (incorrect choice) as well as the waiting time in cases where the immediate reward was chosen. Subjects received 12 trials, with the side on which the immediate reward was presented counterbalanced.

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.

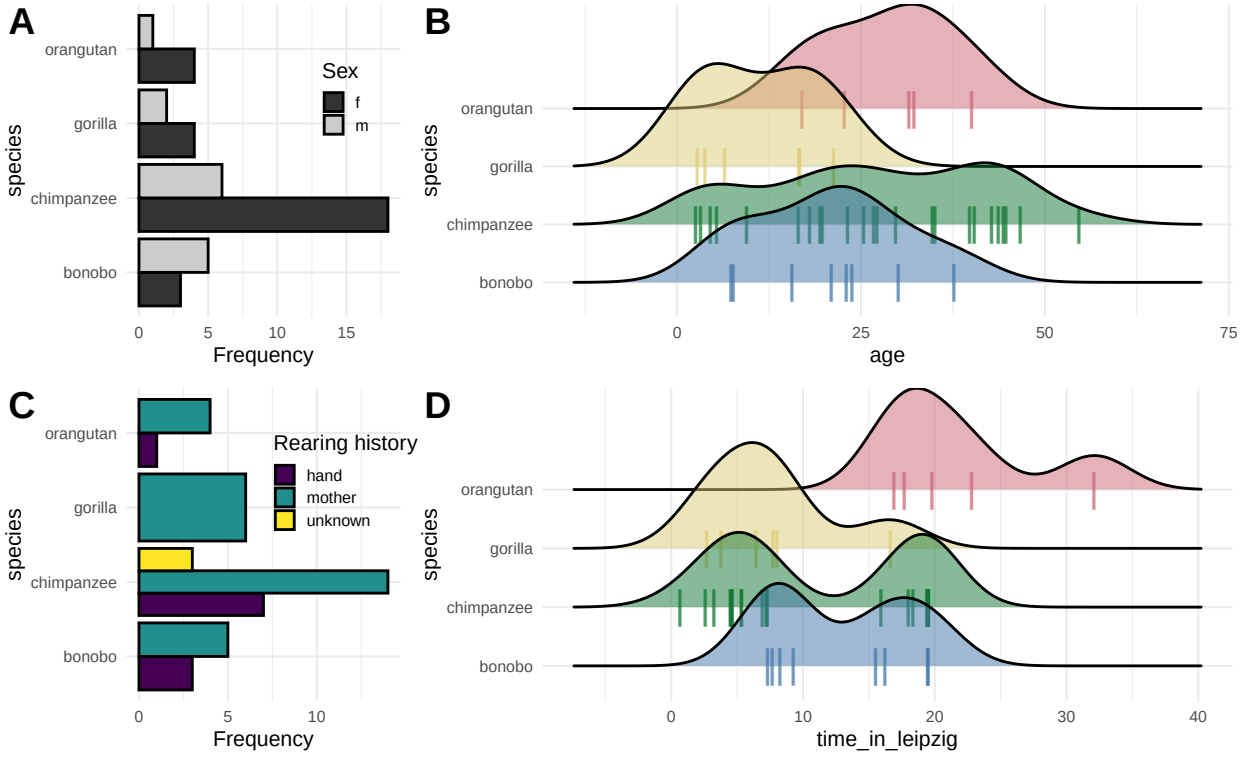

Figure S3: Stable individual characteristics. A) participant sex, B) age distribution by species, C) rearing history, D) time lived in leipzig by species.

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 & 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, 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 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.

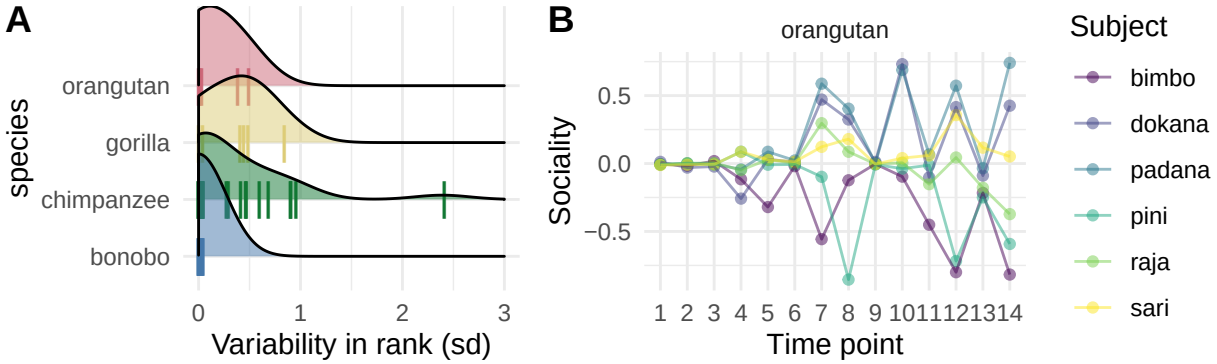

Figure S4: Variable individual characteristics. A) variability in rank (caretaker ratings) for each subject and species, B) sociality estimates for orangutans based on Multiple Membership Relations Model.

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.

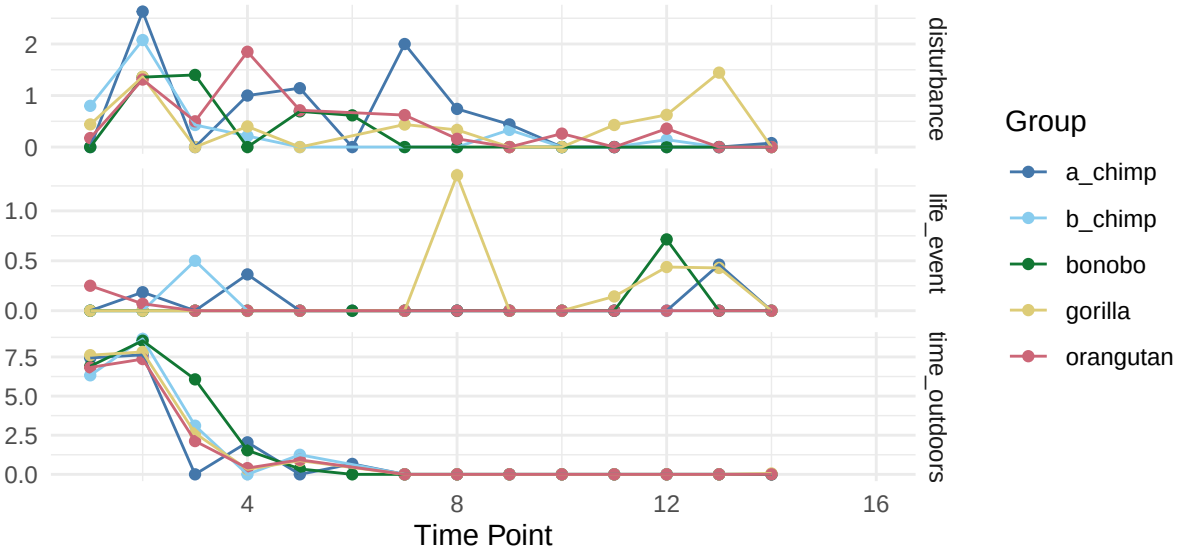

Figure S5: Variation in group life related measures across groups and time points.

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 tasks as as well as the relations between them. We used structural equation modeling [Jana: citation?] 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 – we lay out these assumptions in the text below. The Appendix includes simulations that show that these assumptions were justified.

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 is a variable selection problem (selecting a subset of important predictors from a larger pool) and we used Projection Prediction Inference for this (Piironen, Paasiniemi, & Vehtari, 2018).

Structural equation modelling

Structural equation models can be used to assess latent variables (constructs) using one or more observed variables. These latent variables can be combined in a structural model that imputes relations between them. We used the data from each time point as observed variables to estimate a latent construct for each task. Due to the small sample size, we could not combine latent variables in a structured model or use predictors to explain individual differences in these latent variables. Instead, we assessed relations between tasks by simply correlating latent variables with one another.

We used SEM to estimate states (time varying) and traits (stable over time). In the present context, one can think of traits as a stable psychological ability (e.g. ability to make causal inferences) and states as variable psychological condition (e.g. being attentive). Variation in performance on a given time point can then be partitioned into variance explained by the trait and variance explained by the state. Next we describe the model construction process in more detail.

For each task, 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 with 12 trials per time point this procedure resulted in two test halves assuming 7 possible values (0 to 6 correctly solved trials), for tasks with 8 trials per 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 so sometimes categories had to be collapsed (see descriptions below). 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 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 [Jana: citation?]. For model parsimony, to improve estimation accuracy (see simulation studies) and in order to test for latent mean differences across time, we assume strict measurement invariance. That is, in each model (task), loading parameters are set to 1 at all time points, residual variances are equal to 1, threshold parameters (i.e. trait level necessary to respond above threshold with 0.50 probability) are set invariant across time points and variances of latent state residual factors 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 Model (LSM), which estimates a latent state for each time point based on the two test halves. As

such, it does not assume an underlying trait. Stability is only indirectly assessed via correlations between states. This model first and foremost served as a baseline to see if the data supports a SEM approach.

Second, we fit a Latent State Trait Model (LSTM). This model estimates time point specific states, but also a time-invariant trait. With this model, we can partition the variance in performance into stable (trait) and variable (state) components.

Finally, we fit an LSTM with autoregressive effects (LST-AR). In addition to the LSTM architecture, this model assumes that the states variance at one time point can be used to predict the state variance in the next time point. Thereby it captures the idea that measurements that are closer in time are more likely to be more similar. This allows us to look at longitudinal trends in the state variance.

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

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

At each time point t, a latent state variable S_t, underlying the two observed indicators Y_{1t} and Y_{2t} 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.

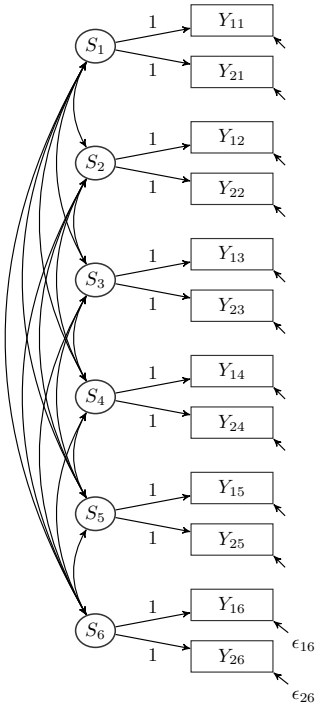

Figure S6: Latent State model for two indicators and six measurement time points.

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

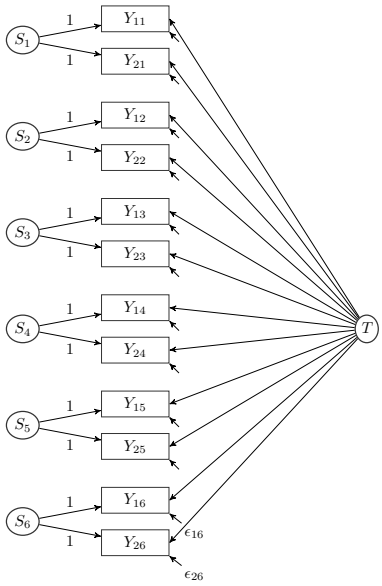

Figure S7: Latent State Trait model for two indicators and six measurement time points.

parameterization; Graded response model). The model is depicted for 6 measurement time points in Figure S7.

AS noted above, we assume strong measurement invariance. As a consequence, the specified LST model (without autoregressive effects) 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.

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 not explained by the trait).

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

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 models with autoregressive effects (LST-AR) This model is described in more detail in Eid, Holtmann, Santangelo, & Ebner-Priemer (2017). The model is depicted for 6 measurement time points in Figure S8.

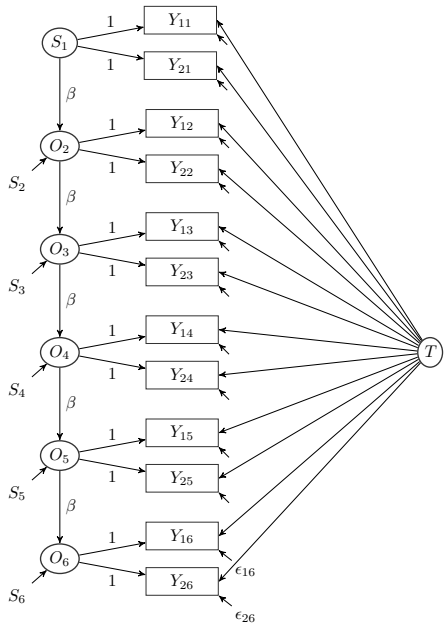

Figure S8: Latent State Trait model with autoregressive effects for two indicators and six measurement time points.

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 cannot be explained by states at previous measurement time points. We make the same assumptions about measurement invariance as in the LST model.

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 (see simulation studies in the Appendix). 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 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 fit was evaluated by computing Posterior Predicted P-values (PPP). The PPP is computed via the following steps: For a given set of parameters (MCMC iterations) a new data set is generated based on the model and those parameters. Then a discrepancy function (e.g. 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 larger, 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 [Jana: sind das die werte die via die discrepancy function berechnet werden? oder wie werden die berechnet?].

For each model, we also report the threshold parameters. The Graded Response Model assumes that the different categories of responses (i.e. the number of correct trials per test half) form an ordered scale. Which category and individual scores, depends om 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 correspond to the level of the latent ability necessary to respond above threshold with 0.50 probability.

Projection predictive inference

The goal of this analysis was to select the predictor variables that are important to predict performance in the different cognitive tasks over time. This constitutes a variable selection problem, for which a range

of different methods are available. We chose to use projection predictive inference because it provides an excellent trade-off between model complexity and accuracy (Piironen & Vehtari, 2017), especially when the goal is to identify **all relevant** predictors (Pavone, Piironen, Bürkner, & Vehtari, 2020).

The projection predictive approach was developed by Piironen, Paasiniemi, & Vehtari (2018) and is used to select a minimal subset of predictors to build a predictive model for cognitive performance. Projective selection can be viewed as a two-step process. The first step is to build the best (exhaustive) predictive model possible which is called the *reference model*. The reference model is a Bayesian multilevel regression model (repeated measurements nested in apes) which includes all predictors recorded in this study. The goal of the second step, is to replace the posterior distribution of the reference model with a simpler distribution. This is by done via a forward step-wise addition of predictors while recording the Kullback-Leibler divergence from the reference model to the projected model at each step. The result is a list containing the best model for each number of predictors.

[for Benedikt: - describe in more detail what the result of the search looks like. How do we come to judge the importance of the different predictors? - describe how the final decision is made and based on which metric]

We implemented this method using the R package `projpred` ("Projpred," n.d.), separately to the data from each of the cognitive tasks. We excluded the switching task because it did not produce any meaningful individual differences.

Results

Stability and Reliability

As mentioned above, we fit three different SEM to the data from each task. Each model offers a different perspective on how stable and reliable performance is. We report the results starting with the LS model, followed by the LST model and finally the LST-AR model.

In the LS models, we can look at stability by comparing the latent means estimated for each time point to see if they differ substantially from one another. For reliability, we can look at the correlations between the latent state estimates for the different time points.

For LST models, we can assess stability by looking at consistency and occasion specificity. A high level of consistency means that a large portion of the variation observed in performance at the different time points can be traced back to variation in the overarching trait. High levels of occasion specificity means the inverse, namely that large portions of the variation in performance is explained by variation in the (residual) state - that is, the variation not explained by the trait. We can quantify reliability as the extend to which our measure of performance is error free, which means, the extend to which performance is explained by variation in the trait and state combined.

LST-AR models use the same metrics as the LST models but in addition they allow us to quantify the temporal predictability of performance based on previous time points. This is captured in the term predictability and quantifies how much of the variation in performance can be explained by the variation in the state at the previous time point.

We ran the same models for the data from phase 1 and phase 2. We first report the results for each task separately for the two phases and then compare how they differ between phases. All models showed acceptable fit indices (see Table S1). The threshold parameters for each model are shown in (see Table S2).

Phase 1

To get an overview of the results, we first visualized he data. Figure S9 shows performance at the different time points. From a group level perspective, we can say that performance was consistently above chance

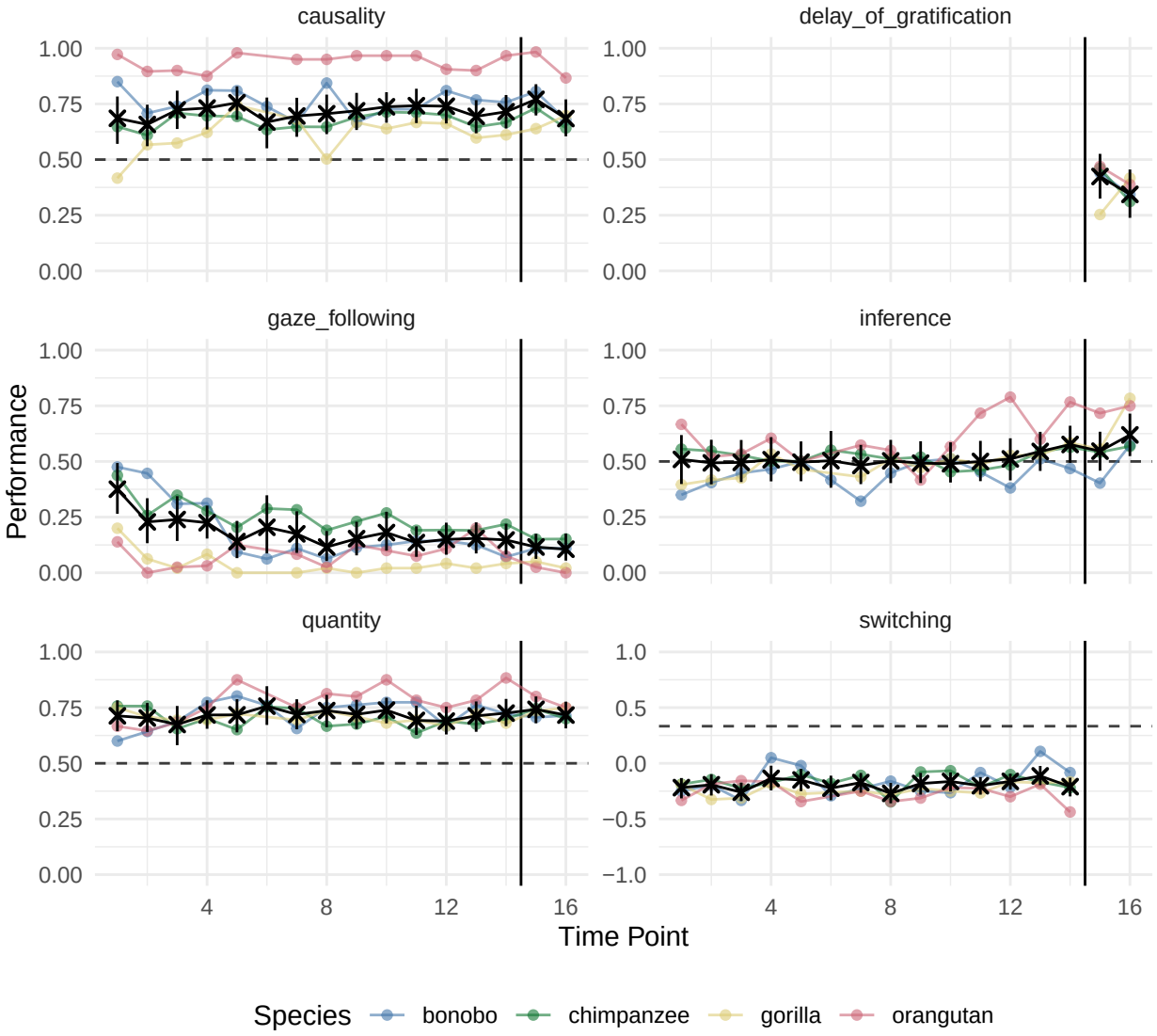

Figure S9: Results from the five cognitive tasks across time points. Black crosses show mean performance at each time point across species (with 95% CI). Colored dots show mean performance by species. Dashed line shows the chance level whenever applicable. The vertical back line marks the transition between phase 1 and 2.

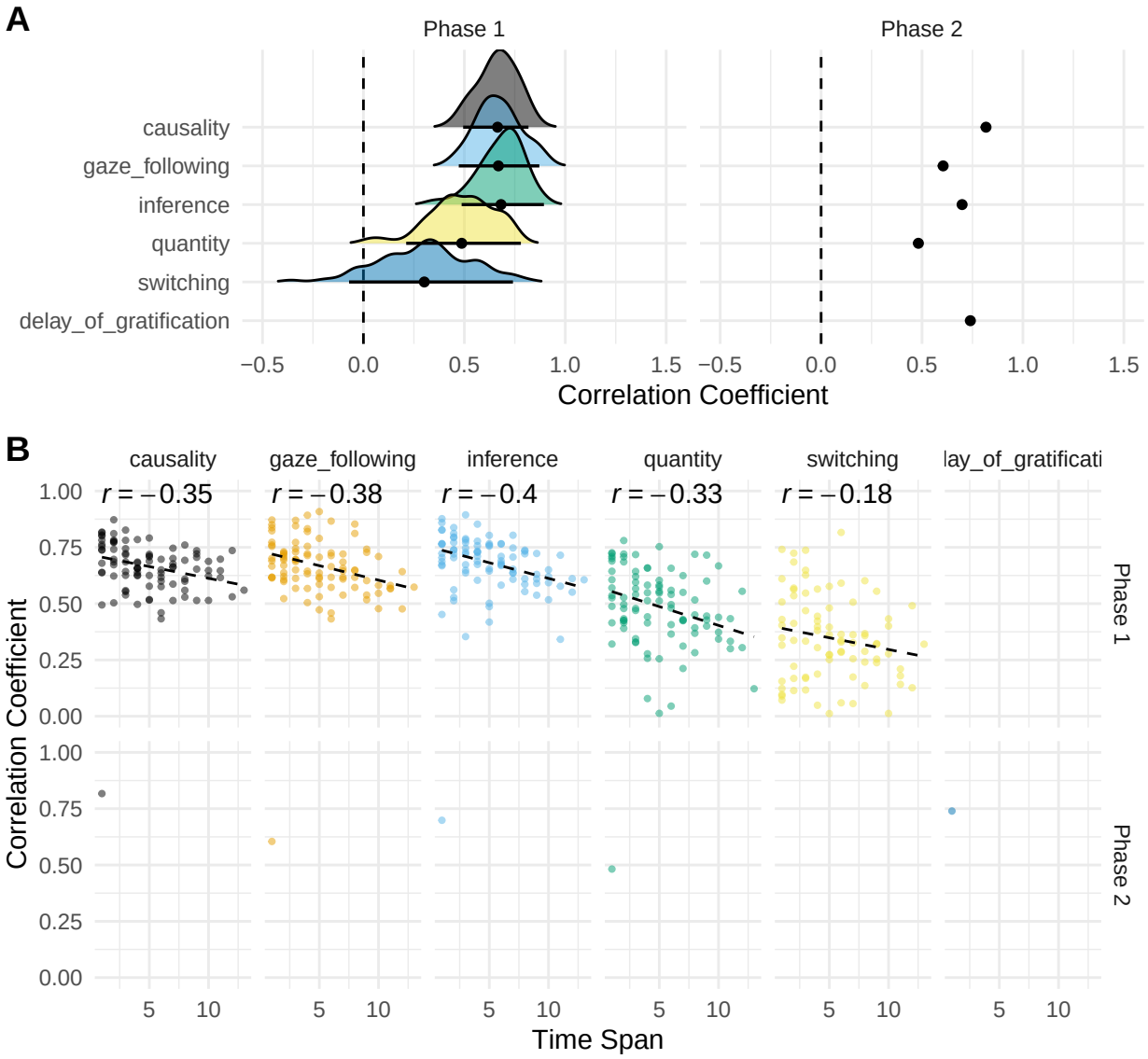

Figure S10: (A) Distribution of correlations between time points for each task. Dots represent the mean of the distribution with 95% HDI. Numbers denote mean and 95% HDI. (B) Correlations between re-test reliability and time span (in time points) between the testing time points.

Table S1: Model fit indices

Task	Model	PPP	Chi 95% CI
causality	LSM	0.242	-74.402;161.09
	LSTM	0.224	-72.051;161.09
	LSTM-AR	0.262	-80.04;156.563
inference	LSM	0.336	-88.002;137.279
	LSTM	0.145	-48.42;137.279
	LSTM-AR	0.197	-65.369;165.291
gaze_following	LSM	0.535	-124.04;111.499
	LSTM	0.360	-99.091;111.499
	LSTM-AR	0.485	-114.891;126.697
quantitiy	LSM	0.485	-103.635;119.705
	LSTM	0.508	-116.335;119.705
	LSTM-AR	0.520	-116.228;108.538

Note:
LSM = Latent state model
LSTM = Latent state trait model
LSTM-AR = LST model with autoregressive component
PPP = Posterior predictive p-value
Chi 95% CI = 95%CI of difference between predicted and observed chi-square values

Table S2: Threshold parameters

Task	Model	T1	T2	T3	T4	T5	T6
causality	LSM	-2.706	-1.717	-1.080	-0.078	0.915	
	LSTM	-2.892	-1.907	-1.268	-0.270	0.728	
	LSTM-AR	-2.919	-1.923	-1.280	-0.264	0.752	
inference	LSM	-2.795	-1.599	-0.715	0.628	1.444	2.672
	LSTM	-2.874	-1.652	-0.736	0.663	1.522	2.808
	LSTM-AR	-2.935	-1.719	-0.805	0.576	1.431	2.712
gaze_following	LSM	-1.204	0.057	1.163			
	LSTM	0.086	1.402	2.547			
	LSTM-AR	0.244	1.561	2.747			
quantitiy	LSM	-1.364	-0.752	0.356	1.411		
	LSTM	-1.398	-0.802	0.254	1.237		
	LSTM-AR	-1.433	-0.832	0.239	1.241		

Note:
LSM = Latent state model
LSTM = Latent state trait model
LSTM-AR = LST model with autoregressive component
T1-6 = Threshold parameters for response categories

(0.5) in the causal inference and quantity tasks. For gaze following, there is no meaningful chance level. We can note, however, that group level performance never went down to zero, which would be expected if apes did not pay attention to the experimenter's gaze. The performance score in the switching task was largely negative, suggesting no successful switching between the two phases.

For an idea of the stability of individual differences, we correlated performance at the different time points for each task (all possible combinations of time points). Figure S10A visualizes the distribution of raw correlations between the different time points and S10B shows the relation between re-test correlations and the time span between time points. Correlations between time points were large and clearly different from zero for quantity, inference and gaze following. For quantity, this distribution was wider and closer to zero, but still clearly positive. For switching, the distribution was even wider and substantially overlapped with zero. For all tasks, correlations between time points were lower for time points that were further apart [Manuel: cite jana uher study].

Given this pattern, we excluded the switching task from further analysis. There were three main reasons for this. First, group level scores were constantly negative and performance in the feature trials always overlapped with chance. This suggests that, as a group, apes did not successfully switch strategies (see S9). Second, the correlations between the different measurement time points were low, suggesting no systematic individual differences (see S10). Third, the dependent variable (i.e. the score calculated based on performance in the two phases) had a different level of measurement compared to the other tasks. That is, there was only a single score to represent at each time point. All other tasks had multiple trials. This was especially problematic in the context of structural equation modeling (see above). For these reasons we also replaced the switching task with the delay of gratification task in phase 2.

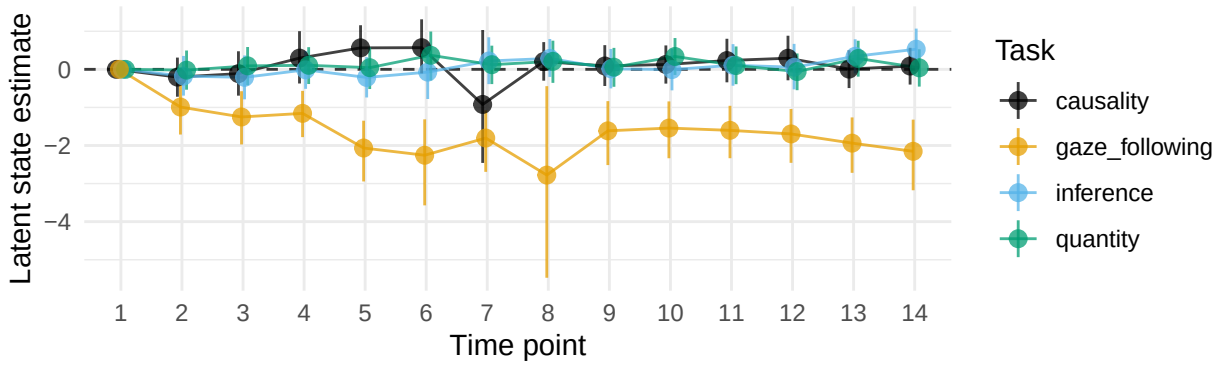

Figure S11: Latent means for LSM with 95% CI. Means at time point 1 are set to 0.

Causal inference To fit the models, the response categories of 0 or 1 solved trial 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 3 and 11 as well as test-half 1 at time points 4 and 12 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 respective state factors based on the respective other test half. At time point 7 thresholds of both test-halves could not be set measurement invariant across time (due to a divergent number of observed categories). Latent mean differences with respect to the latent state variable at time point 7 should therefore be interpreted with caution.

Figure **??** visualizes the latent state means estimated in the LS model. None of the means differed significantly from zero, suggesting that there was no systematic change over time. Figure **??** gives the correlations between the latent states for the different time points. Correlations were generally high, indicating stable and reliable individual differences.

In the LSTM, the consistency (i.e., stability) coefficient was estimated to be around .903. This means that around 90% of true inter-individual differences are attributable to stable differences between individuals,

while approximately 10% are due to variance in time-point specific deviations from the stable trait. Reliability was high with an estimate of .725 (see Figure S13).

Figure S14 shows the parameters from the LSTM-AR for three time points (2, 3 and 14). Around 82.3% of true interindividual differences at time point 14 go back to stable trait differences, around 10.6% of the interindividual differences can be explained by carry-over effects from previous time points (i.e. inertia in the within-person process) and only 6.1% of the variance is due to time-specific variance between persons, that is, variance in the time-specific true scores from the stable trait level that could not be predicted by the autoregressive process.

In sum, all models converge on the conclusion that group- and individual-level performance was highly stable over time. Figure shows that performance is above chance

Inference by exclusion Thresholds could not be set equal for indicator 2 at time point 6 as well as indicator 1 at time points 7 and 14, due to a different number of observed categories for the respective indicator and time-point combination. Latent means can still be compared across time for the respective state factors based on the other indicator.

None of the latent means estimated in the LS model differed from zero (Figure **??**). Correlations between latent states was generally high across time points (Figure **??**).

In the LSTM, consistency was estimated to be around .859 – around 86% of true inter-individual differences were attributable to stable differences between individuals. Approximately 14% were due to variance in time-point specific deviations from the stable trait. Reliability was high with an estimate of .815 (see Figure S13).

According to the LSTM-AR, around 79.4% of true interindividual differences at time point 14 went back to stable trait differences and around 8.3% of the interindividual differences can be explained by carry-over effects from previous time points. Around 11.3% of the variance was due to time-specific variance between individuals.

Taken together, we see a similar pattern as for the causal inference task: Performance was very stable on a group level and so were the differences between individuals. interesting performance at chance but individual differences stabel.

Phase 2

Comparison between phases

Relations between tasks

Phase 1

Phase 2

Comparison between phases

Predictability

Previous section showed that performance was stable and largely explained by stable / trait differences. This limits how much the predictors can actually explain - and as we will see the most important predictor in all cases was a random intercept term for each individual.

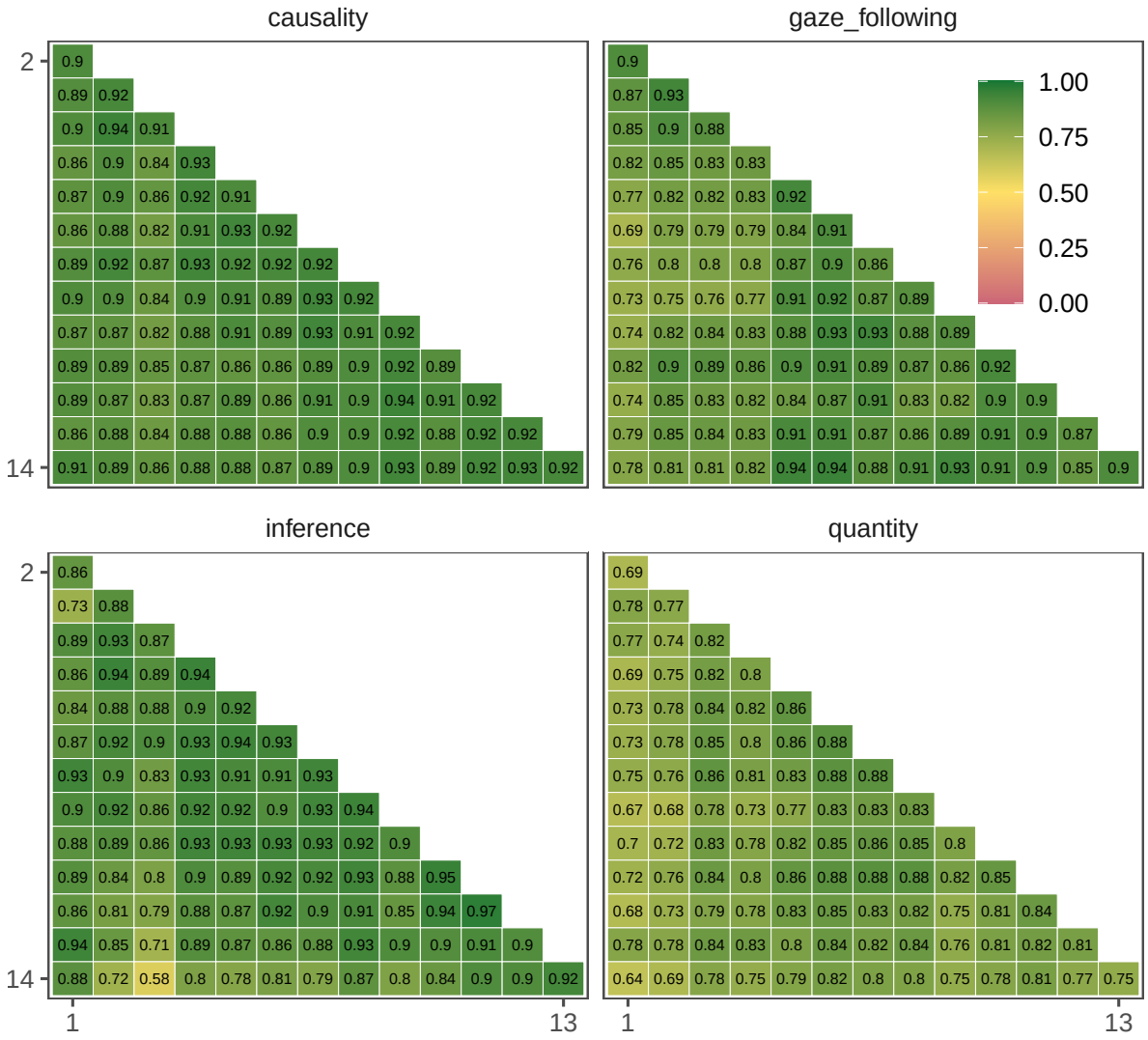

Figure S12: Correlations between latent state variables based on LSM for the different tasks.

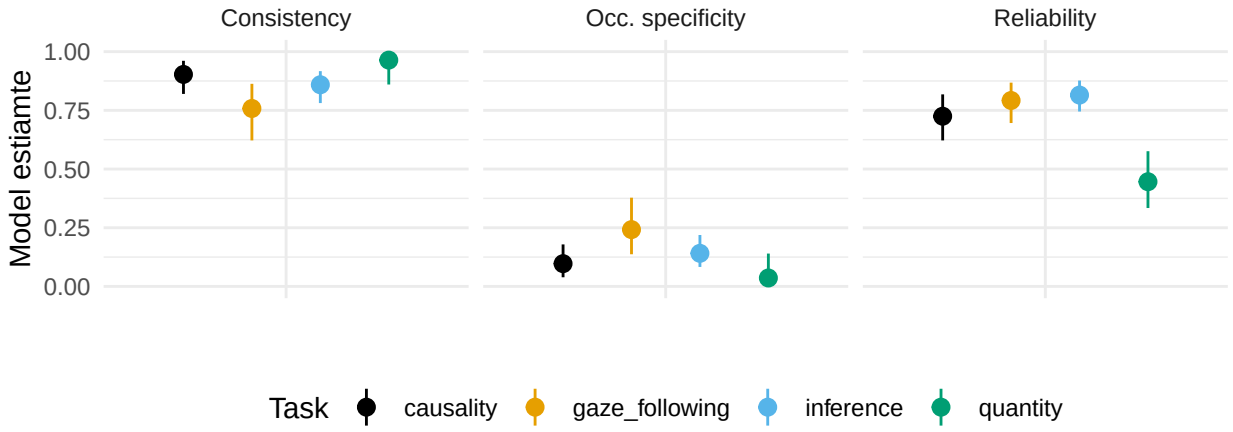

Figure S13: Model parameters (with 95% CI) from LSTM for the four tasks.

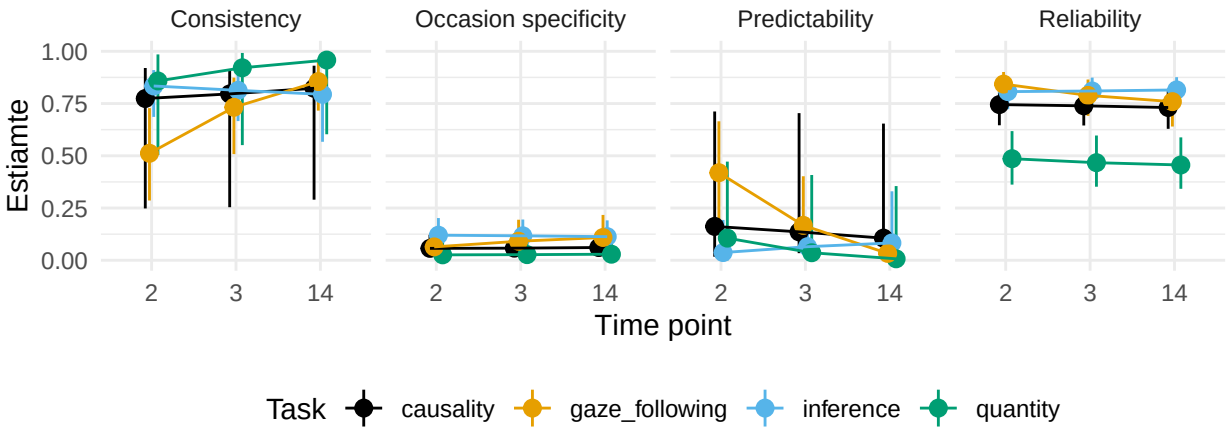

Figure S14: Model parameters (with 95% CI) from LSTM-AR for the four tasks.

Phase 1

Phase 2

Comparison between phases

Summary

Appendix

SEM Simulations

Brauer, J., Call, J., & Tomasello, M. (2005). All great ape species follow gaze to distant locations and around barriers. *Journal of Comparative Psychology*, *119*(2), 145.

Call, J. (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, 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.

Hanus, D., & Call, J. (2007). Discrete quantity judgments in the great apes (pan paniscus, pan troglodytes, gorilla gorilla, pongo pygmaeus): The effect of presenting whole sets versus item-by-item. *Journal of Comparative Psychology*, *121*(3), 241.

Haun, D. B., Call, J., Janzen, G., & Levinson, S. C. (2006). Evolutionary psychology of spatial representations in the hominidae. *Current Biology*, *16*(17), 1736–1740.

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

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

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

Piironen, J., Paasiniemi, M., & Vehtari, A. (2018). Projective inference in high-dimensional problems: Prediction and feature selection. *arXiv Preprint arXiv:1810.02406*.

Piironen, J., & Vehtari, A. (2017). Comparison of bayesian predictive methods for model selection. *Statistics and Computing*, *27*(3), 711–735.

Projpred: Projection predictive feature selection. (n.d.). Retrieved from https://mc-stan.org/projpred

Rosati, A. G., Stevens, J. R., Hare, B., & Hauser, M. D. (2007). The evolutionary origins of human patience: Temporal preferences in chimpanzees, bonobos, and human adults. *Current Biology*, *17*(19), 1663–1668.

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