

tbd...

Supplementary material

tbd...

## Contents

<b>Overview</b>	<b>1</b>
<b>Methods</b>	<b>2</b>
Participants . . . . .	2
Setup . . . . .	3
Tasks . . . . .	3
Data collection . . . . .	5
Predictors . . . . .	6
<b>Analytical framework</b>	<b>10</b>
Structural equation modeling . . . . .	10
Projection predictive inference . . . . .	16
<b>Results</b>	<b>17</b>
Stability and Reliability . . . . .	17
Relations between tasks . . . . .	26
Predictability . . . . .	28
<b>Summary</b>	<b>33</b>
<b>References</b>	<b>33</b>
<b>Appendix</b>	<b>34</b>
SEM Simulations . . . . .	34

## Overview

This document gives a detailed overview of the methods used in the study “...tbd...” First, we give an overview of our great ape participants. Next we describe the general setup and the experimental tasks that were used. 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 describe the two parts of our analytical framework: Structural Equation Modeling to investigate stability and reliability of cognitive performance and Projection Predictive Inference to test the importance of the predictor variables.

We present the results separate for the two phases of data collection. For each phase, we first report results on stability and reliability of performance within each task and then we investigate relations between performance in the different tasks. Finally, we report how the different predictors related to performance in the different tasks.

The appendix contains results from simulation studies we conducted to investigate the performance of the employed Structural Equation Models under the sample sizes given in the present dataset.

## Methods

### Participants

A total of 44 great apes participated at least once in one of the tasks. This included 8 Bonobos (3 females, age 7.3 to 38.8), 24 Chimpanzees (18 females, age 2.6 to 55.9), 6 Gorillas (4 females, age 2.7 to 22.5), and 6 Orangutans (4 females, age 17 to 41.1). The sample size at the different time points ranged from 1 to 18 for the different species. Figure 1 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 participated in cognitive research on a regular basis. Many of them had 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 (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.

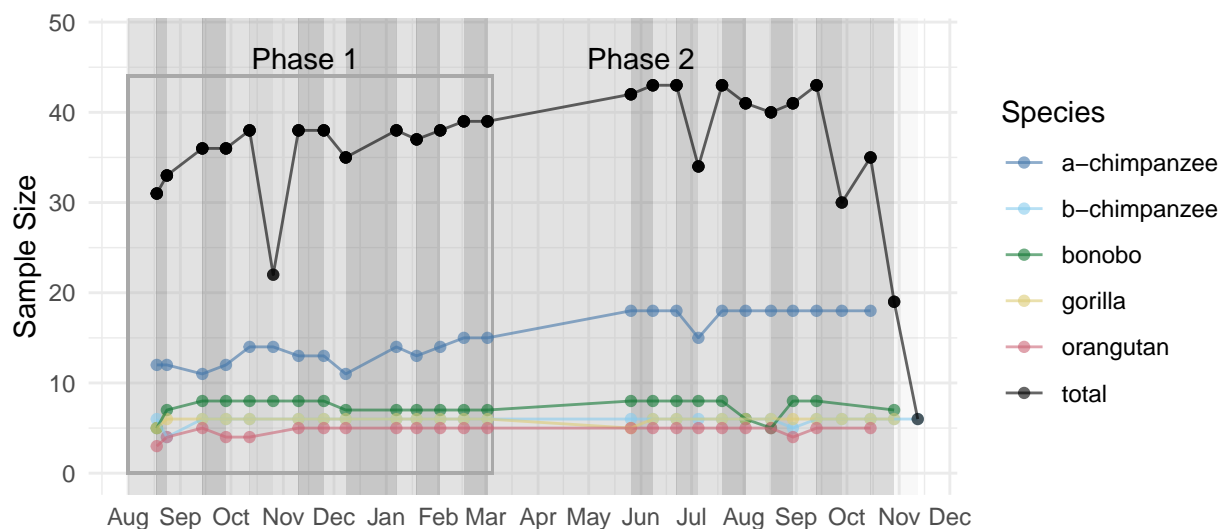


Figure 1: Sample size by species across the different time points. Time point specific predictor variables were collected during the time between two time points (shaded regions) to predict the next.

## 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 2).

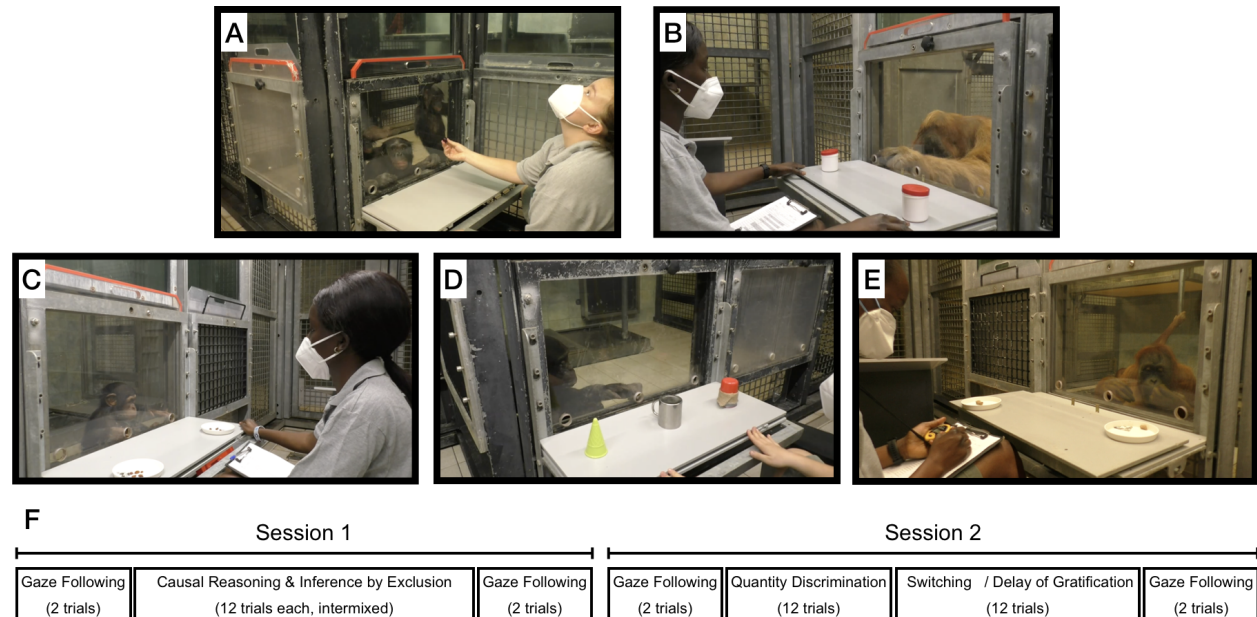


Figure 2: Setup used for the six tasks. A) Gaze following. B) Causal reasoning and inference by exclusion. C) Quantity discrimination. D) Switching. E) Delay of gratification. F) Order of task presentation and trial numbers

## 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/](#).

### 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 2C). The experimenter looked to the ceiling, no object of particular interest was placed there. After 10s, the experimenter looked

down again, handed over the food and the trial ended. We coded whether the participant looked up during the 10s interval. Apes received eight gaze following trials.

We assume that participants look up because they assume that the experimenter’s attention is focused on a potentially noteworthy object.

### **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 2A). The experimenter covered the table with the occluder, retrieved a piece of food, showed it to the ape, and hid it in one of 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 of the food was counterbalanced; six times in the right cup and six times in the left. Causal inference trials were intermixed with inference by exclusion trials (see below).

We assume that apes locate the food by reasoning that the food – a solid object – causes the rattling sound and therefore must 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: six times the left cup was baited, six 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.

### **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 2B). The experimenter covered the plates with the occluder and placed five small food pieces on one plate and seven 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, six with the larger quantity on the right and six on the left (order counterbalanced).

We assume that apes identify the larger of the two food amounts based on discrete quantity estimation.

### **Switching**

This task was modeled after Haun, Call, Janzen, & Levinson (2006). Three differently looking cups (silver metal cup with handle, green plastic ice cone, red cup without handle - Figure 2D) 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 subsequent 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:  $[\text{proportion correct place}] - (1 - [\text{proportion correct feature}])$ . Positive values in this score mean that participants could quickly switch from choosing based on location to choosing based on feature. High negative values suggest that participants did not or hardly switch strategies.

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

The switching task was only used in Phase 1. It did not produce meaningful results (see results for Phase 1 in Stability and Reliability below) and for Phase 2 we therefore replaced it with a delay of gratification task (see below).

### Delay of gratification

This task replaced the switching task in Phase 2. The procedure 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.

We assume that, in order to choose the larger reward, apes inhibit choosing the immediate smaller reward.

### 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 2F). Session 1 started with two gaze following trials. Next was a pseudo randomized mix of causal inference and inference by exclusion trials 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 two gaze following trials. Session 2 also started with two gaze following trials, followed by quantity discrimination and switching (Phase 1) or Delay of Gratification (Phase 2). Finally, there were again two gaze following trials. By 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 and time points. The two sessions were usually spread out across two adjacent days. For the larger chimpanzee group, they were sometimes spread out across four 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 some intervals were longer or shorter. Figure 1 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 1). Phase 2 started on May 26th, 2021 and 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 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. 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 3 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`.

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

### 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 Figure 4A) across time points, however, there was some variation. Variable name in model: `rel_rank`.

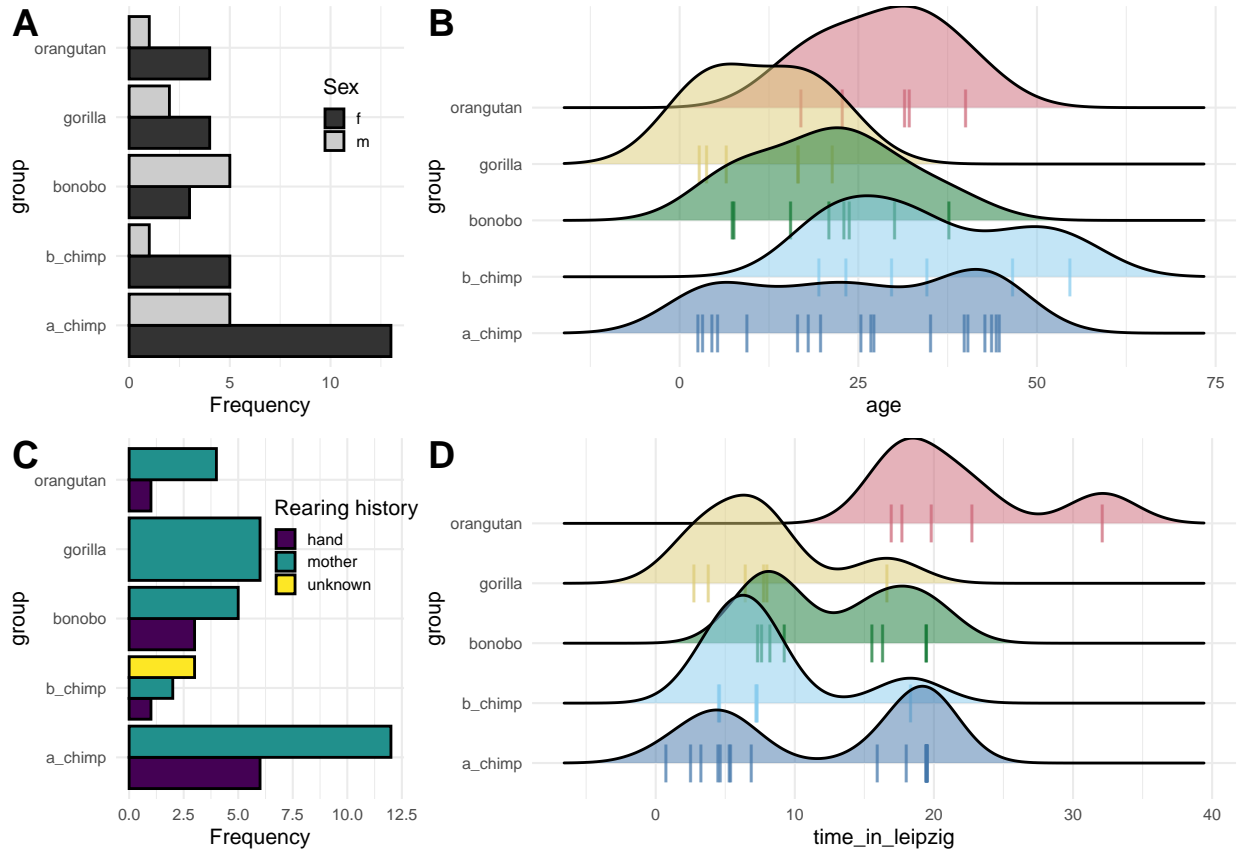


Figure 3: Stable individual characteristics. A) participant sex, B) age distribution by species, C) rearing history, D) time lived in leipzig by species.

**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: `sick_severity`.

**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 were within arms reach. Research assistants used a tablet to record their observations with the behavioral coding software ZooMonitor (Wark et al., 2019).

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. Figure 4B visualizes the sociality measures for one group across the different time points. Variable name in model: `sociality`.

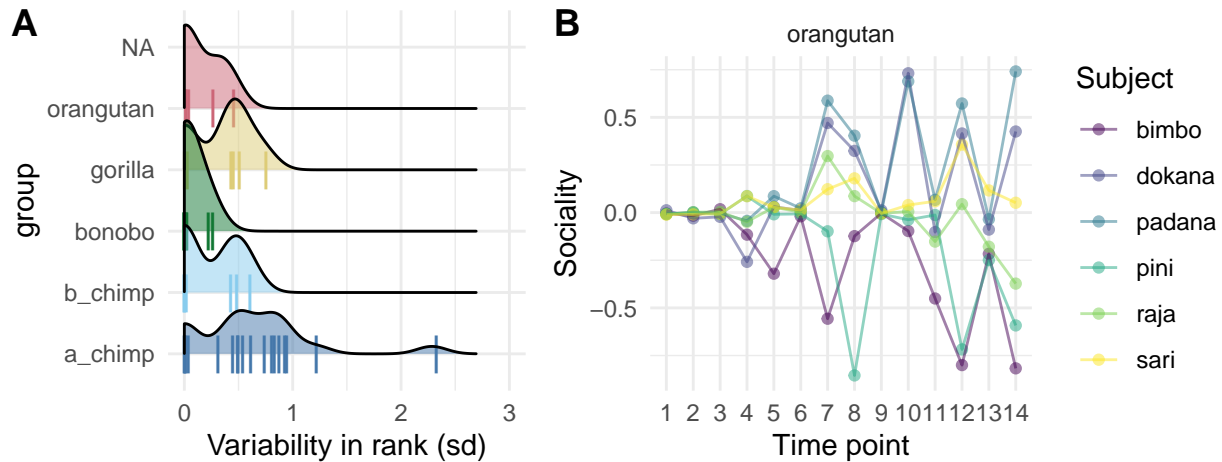


Figure 4: Variable individual characteristics. A) variability in rank (caretaker ratings) for each subject and species, B) sociality estimates for orangutans based on Multiple Membership Relations Model.

## Group life

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. Figure 5 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_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: `dist_mean`.

**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: `le_mean`.

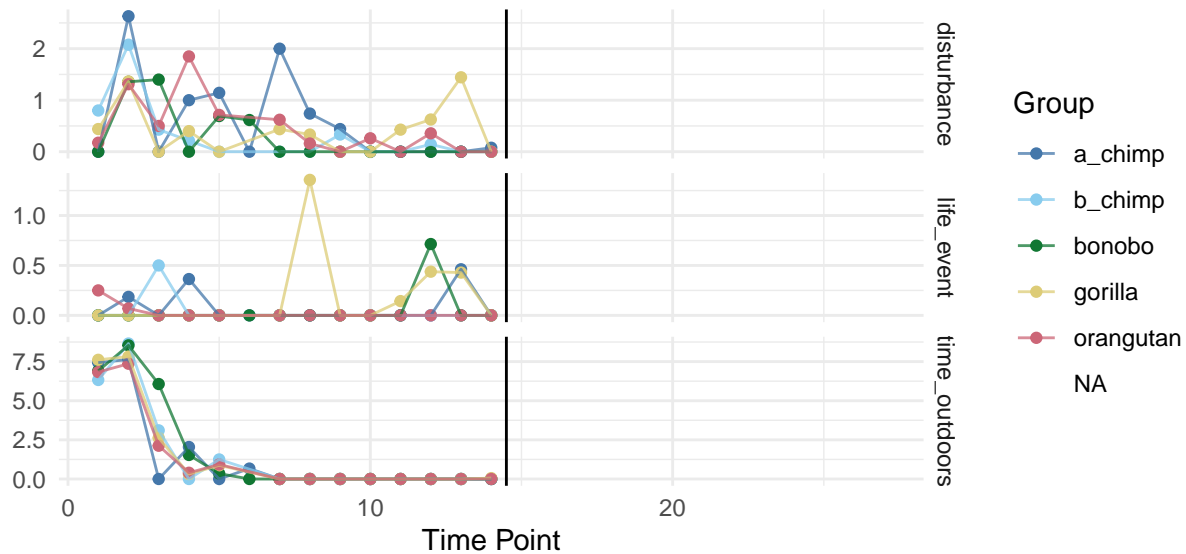


Figure 5: Variation in group life related measures across groups and time points.

## 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. Variable name in model: `observer`.

**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: `test_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: `test_tp`.

## 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 stable performance in a given task was on a group-level, how stable individual differences were, 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 appendix we present results from a small simulation study which show that parameters in the employed SEMs are 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 is a variable selection problem (selecting a subset of variables from a larger pool) and we used *Projection Prediction 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 (on a stable as well as the time-point specific level). 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.

We used SEM to estimate traits (stable over time) and states (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 states as time-specific, 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 explained by the trait, variance explained by the situation or individual-situation interactions, and measurement error. Because the latent variables are estimated on multiple indicators, 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 several identical trials per individual per task. Using the individual trials (8 to 12 dichotomous items, depending on the task) as indicators for a latent ability factor per time point (i.e., assuming a Rasch model per time point) in a longitudinal SEM resulted in estimation problems due to many empty cells in the bivariate distributions across time and / or tasks. Therefore we decided to model sum scores of the repeated trials, 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 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 reliability). Due to only few observed categories 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 (see simulation studies) and in order to test for latent mean differences across time, we assume strict factorial (or measurement) invariance across time [Manuel: passende Quellen wären hier: Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, 58, 525–543. doi: 10.1007/BF02294825 Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered categorical measures. *Multivariate Behavioral Research*, 39, 479–515. doi: 10.1207/S15327906MBR39034.]. 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.

[Manuel: den Satz zu equal state residual variances hab ich nach unten verschoben, da nur für das LST und nicht das LS Modell gültig. ].

## 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. Stability of group 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. [Manuel: hab den Absatz zu less restrictive than LST zum LST Modell verschoben, da dort im direkten Vergleich besser verständlich]

Second, we fit a Latent State Trait Model (LSTM). In LST models, true inter-individual differences are decomposed into a (time-invariant) latent trait variable and time-specific deviations of the true score from the stable trait (state residual variable). The model allows 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.

Finally, we fit an LSTM with autoregressive effects (LST-AR). In addition to the LSTM architecture, this model assumes that the state residual variables at one time point can be used to predict the true score at the subsequent time point. As such, it captures the idea that measurements that are closer in time are likely to be more highly correlated and quantifies potential carry-over effects from one time point to the next. 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  $S_{it} = 0, \dots, k_{it} - 1$  the relation between  $Y_{it}$  and  $Y_{it}^*$  is described by:

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

where  $\tau_{sit}$  denote threshold parameters [Manuel: Referenz hier einfügen: Muthen, B. O. (1984). A general structural equation model with dichotomous, ordered categorical and continuous latent variable indicators. *Psychometrika*, 49, 115-132.]

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  $\tau_{sit}$  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  $\epsilon_{it}$  (see, for instance, [Manuel: Eid, M., & Kutscher, T. (2014). Statistical models for analyzing stability and change in happiness. In K. M. Sheldon & R. E. Lucas (Eds.), *Stability of happiness: Theories and evidence on whether happiness can change* (pp. 261–297). London: Elsevier]):

$$Y_{it}^* = S_t + \epsilon_{it} \quad (2)$$

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.

[Manuel: Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, 52, 393–408. doi: 10.1007/BF02294363; ]

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 Figure 6.

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  $\tau_{sit}$  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:

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

[Manuel: das bringt vermutlich deine Nummerierung der Gleichungen durcheinander, entschuldige, hab das label erstmal rausgenommen.]

**Latent State Trait (LST) models** In LST models, the latent state variable is further decomposed into a latent trait variable  $T$  and a latent state residual variable  $\zeta_t$ , resulting in the following measurement equation for parcel  $i$  at time point  $t$  [Manuel: bitte Quelle einfügen: Steyer, R., Schmitt, M., & Eid, M. (1999). Latent state-trait theory and research in personality and individual differences. *European Journal of Personality*, 13, 389–408.]:

$$Y_{it}^* = T + \zeta_t + \epsilon_{it} \quad (4)$$

Here,  $T$  is a stable (time-invariant) latent trait variable, capturing stable interindividual differences between individuals. The state residual variable  $\zeta_t$  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.  $\epsilon_{it}$  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 Figure 7.

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.

[Manuel: die folgenden zwei sätze würde ich rausnehmen, da hier eigentlich irrelevant: LST models as single-level models are estimated to test whether measurement invariance assumptions across time can be

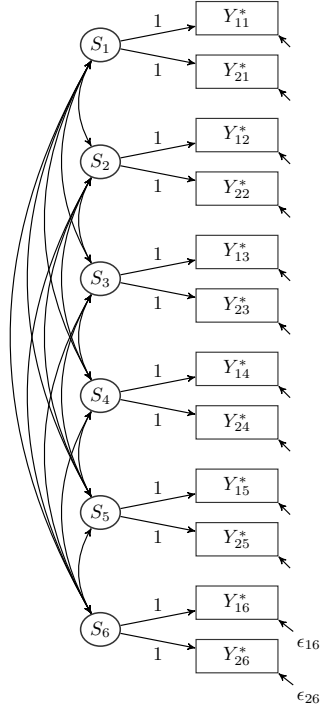


Figure 6: Latent State model for two indicators and six measurement time points.

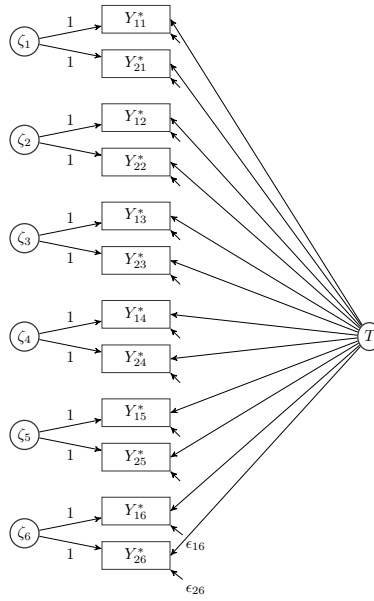


Figure 7: Latent State Trait model for two indicators and six measurement time points.

reasonably assumed. Once measurement invariance can be established, the models can alternatively be estimated as multilevel SEMs.]

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.

$$Con(Y_{it}^*) = \frac{Var(T)}{Var(T) + Var(\zeta_t)} \quad (5)$$

**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).

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

As state residual variances  $Var(\zeta_t)$  were 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 six measurement time points in Figure 8.

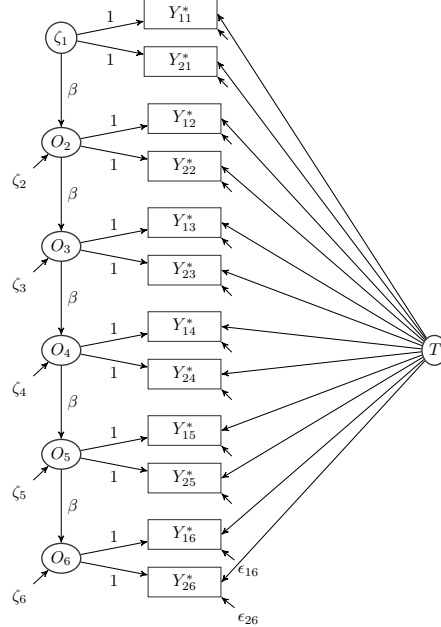


Figure 8: Latent State Trait model with autoregressive effects for two indicators and six measurement time points.

Measurement equation for parcel  $i$  at time point  $t$ :

$$Y_{it}^* = T + O_t + \epsilon_{it} \quad (7)$$

where  $T$  is a stable latent trait variable, the occasion-specific variables  $O_t$  capture time-specific deviations of the respective true score from the stable trait  $T$ , and  $\epsilon_{it}$  is a measurement error variable, with  $\epsilon_{it} \sim N(0, 1) \quad \forall i, t$ .  $O_t$  is assumed to follow an autoregressive process of order 1 across time (within subjects), that is:

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

where the latent state residual variables  $\zeta_t$  capture true time-specific inter-individual differences that cannot be explained based on the true score at previous measurement time points. We make the same assumptions about factorial 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.

$$Con(Y_{it}^*) = \frac{Var(T)}{Var(T) + \beta^2 Var(O_{(t-1)}) + Var(\zeta_t)} \quad (8)$$

**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 (or the stable trait).

$$OS(Y_{it}^*) = \frac{Var(\zeta_t)}{Var(T) + \beta^2 Var(O_{(t-1)}) + Var(\zeta_t)} \quad (9)$$

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.

**Autoregressive predictability** Proportion of true variance that is explained by carry-over effects from previous measurement time points:

$$Pred(Y_{it}^*) = \frac{\beta^2 Var(O_{(t-1)})}{Var(T) + \beta^2 Var(O_{(t-1)}) + Var(\zeta_t)} \quad (10)$$

**Models for combinations of tasks** To investigate associations between cognitive performance in different tasks, the described models were extended to multitrait models. Due to the small sample size, we could not combine all tasks in a single, structured model. Instead, we assessed relations between tasks in pairs.

## Estimation

Models were estimated with MPlus version 8.4 [Manuel: hier müssten wir das Handbuch zitieren: Muthen, L.K. and Muthen, B.O. (1998-2017). Mplus User's Guide. Eighth Edition. Los Angeles, CA: Muthen & Muthen], 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 fit was evaluated by computing Posterior Predicted P-values (PPP). 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 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. [Manuel: wenn die fit function der chi-quadrat wert ist, dann wird geschaut, ob der Wert für die echten Daten kleiner ist als für die replizierten Daten (kleiner = besserer fit), also genau anders herum.] 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?]  
 [Manuel: Ja, die predicted und observed chi-square values werden verglichen und in den PPP umgerechnet wie oben beschrieben, wobei  $PPP = \text{Prob}(\text{chi-square real data} < \text{chi-square replicated data})$ ; die Verteilung der Differenzen aus den beiden ist ein zusätzlicher Indikator, dessen KI die 0 überdecken sollte; hier noch eine ergänzende Beschreibung: 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 & Muthen, 2010).]

## 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 (e.g. Lasso Regression). 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 a minimal subset of predictors that yield a good predictive model (Pavone, Piironen, Bürkner, & Vehtari, 2020).

The predictive projection approach was developed by Piironen et al. (2020). It is used to select a minimal subset of predictors that allow to generate an accurate predictive model for cognitive performance. Projective selection can be viewed as a two-step process. The first step consists of building the best predictive model possible, called the reference model. The reference model is a Bayesian multilevel regression model (repeated measurements nested in apes), including all available predictors. 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 divergence from the reference model to the projected model. The result is a list containing the best model for each number of predictors. The final model is selected by inspecting the mean log-predictive density and/or root-mean squared error. The projected model with the smallest number of predictors that shows similar predictive performance as the reference model is chosen.

We built four different (Bayesian multilevel regression) reference models and ran them through the above predictor selection process for the following four tasks: Gaze following, causal inference, inference by exclusion, and quantity discrimination. The dependent variable for each task was the cognitive performance of the apes or, i.e., how well they executed the tasks with success. Continuous predictors were centered when needed. We transformed the apes rank variable into a relative rank, where a rank with value one depicts a subject with the highest possible rank. For gaze following, we added the predictor `day2`, which simply indicated if the trials were from the second session or the first. 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 convergence of the model; 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. The effective sample size (ESS) is an estimate within a chain for the number of independent samples containing the same amount of information about the dependent variable.

Next, we performed the predictor selection for each reference model separately, thus resulting in four different rankings for the relevant predictors. The predictor selection was executed with the R package **projpred** (“Projpred,” n.d.), which implements the aforementioned predictive projection technique. The predictor relevance ranking is measured by the LOO cross-validated mean log-predictive density (**elpd**) and root-mean-squared error (**rmse**). 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 the most goes into the ranking. 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. Alternatively, one could have used the function **suggest\_size** as a heuristic decision rule to find the optimal submodel. **Suggest\_size** chooses the smallest submodel with an **elpd** within one standard error of the reference model (default rule). The smallest submodel is thus expected to outperform the reference model with at least 16% probability. However, this is not yet possible for the models we fit due to the delay of the random intercept term.

## Results

### Stability and Reliability

As mentioned above, we fit three different SEM to the data from each task. Each model offers a slightly 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.

Within the context of SEM, reliability is defined as the proportion of the variance that is error-free, that is, variance that is explained by the latent variables (state (residual) and/or trait variables). Reliability is estimated based on the correlations between indicators. Because the two indicators corresponded to the two test halves in our case, reliability was equivalent to a split-half reliability estimate.

In the LS models, we can look at the stability of group level performance by comparing the latent means estimated for each time point to see if they differ substantially from one another. To assess the stability of individual difference, we can look at the correlations between the latent state estimates for the different time points.

For LST models, we can assess stability of individual differences by looking at consistency and occasion specificity estimates. 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 state residual – that is, occasion-specific variation not explained by the trait, due to situation and / or person-situation interaction effects.

LST-AR models extend LST models by accounting for potential carry-over effects between temporally adjacent observations. The autoregressive effects thereby capture accumulated situational effects, that is, effects of previously experienced situations on subsequent trait levels (Eid et al., 2017). For instance, a person that experiences a great success at work at time  $t$  may not only show greater job satisfaction as compared to their

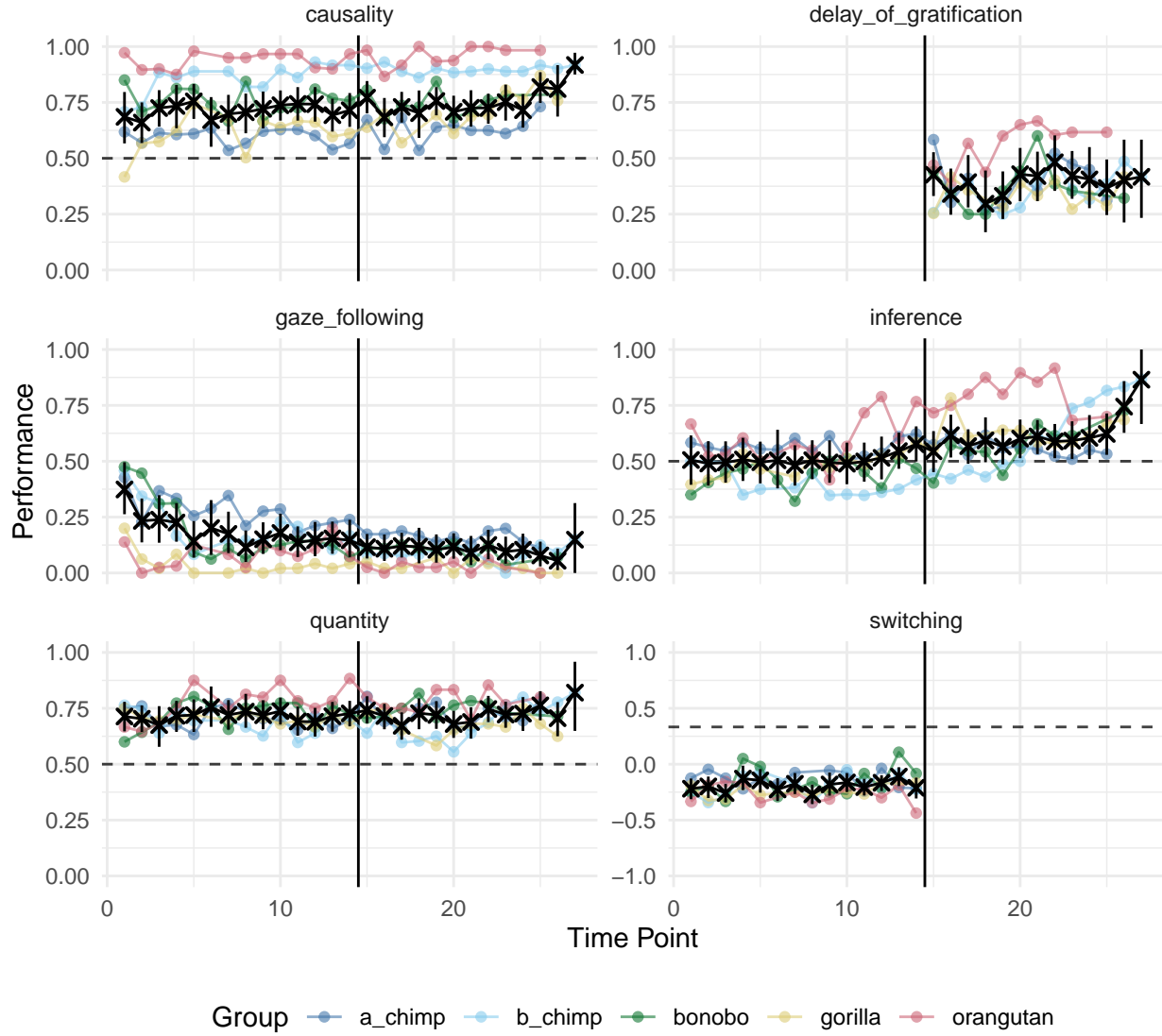


Figure 9: Results from the five cognitive tasks across time points. Black crosses show mean performance at each time point across species (with 95% CI). Colored dots show mean performance by species. Dashed line shows the chance level whenever applicable. The vertical back line marks the transition between phase 1 and 2.

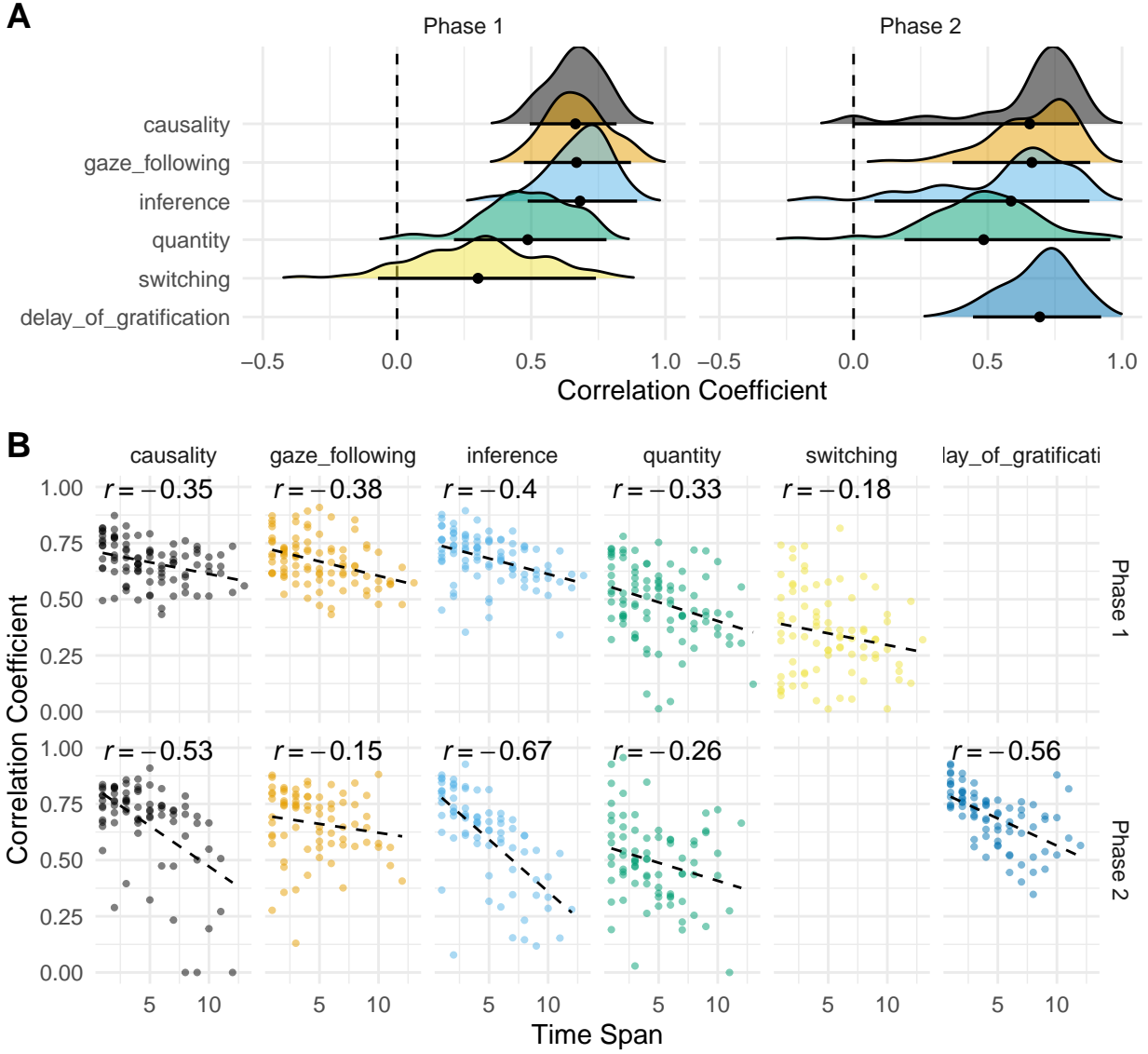


Figure 10: (A) Distribution of correlations between time points for each task. Dots represent the mean of the distribution with 95% HDI. Numbers denote mean and 95% HDI. (B) Correlations between re-test reliability and time span (in time points) between the testing time points.

Table 1: Model fit indices

Task	Model	PPP	Chi 95% CI
causality	LSM	0.242	-74.40 ; 161.09
	LSTM	0.224	-72.05 ; 161.09
	LSTM-AR	0.262	-80.04 ; 156.56
inference	LSM	0.336	-88.00 ; 137.28
	LSTM	0.145	-48.42 ; 137.28
gaze_following	LSTM-AR	0.197	-65.37 ; 165.29
	LSM	0.535	-124.04 ; 111.50
	LSTM	0.360	-99.09 ; 111.50
	LSTM-AR	0.485	-114.89 ; 126.70
quantity	LSM	0.485	-103.64 ; 119.70
	LSTM	0.508	-116.33 ; 119.70
	LSTM-AR	0.520	-116.23 ; 108.54

*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

habitual trait level at that point in time but an increase in job satisfaction that lasts across a prolonged time period (i.e., change in the habitual trait level at time  $t > 1$  due to the positive experience at time  $t$ ). LST-AR models allow us to quantify the temporal predictability of performance based on occasion-specific variance in the previous time points. This is captured in the predictability coefficient and quantifies how much of the variation in performance can be explained by the variation in the occasion-specific variables at the previous time point (note that the names of the coefficients used here slightly differ from those used in Eid et al. (2017)).

[Jana: Hast Du hier ein Beispiel aus der Menschenliteratur was das sein könnte. Es ist ja nicht trait, also nicht komplett stabil, sondern was, dass nur naheliegende Zeitpunkte betrifft. Eine Sache die ich mir denken könnte wäre, z.b. Unaufmerksamkeit aufgrund von Krankheit. Trifft es das?]. [Manuel: Jein. Unaufmerksamkeit sollte vermutlich eher nur zu dem Zeitpunkt wirken an dem sie vorhanden ist, und nicht darüber hinaus (es sei denn sie demotiviert wegen damit verbundener höherer Trefferquote etc). Ehrlich gesagt ist die Theorie dahinter leider etwas komplizierter als wir sie hier darstellen. Ich habe den Absatz dementsprechend umformuliert und ein Bsp eingefügt. Ich hoffe es ist so verständlich?]

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 1). The threshold parameters for each model are shown in (see Table 2).

## Phase 1

To get an overview of the results, we first visualized the data. Figure 9 shows performance at the different time points. From a group-level perspective, we can say that performance was consistently above chance (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.

Table 2: 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

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). Figure 10A visualizes the distribution of raw correlations between the different time points and 10B 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 tended to be lower for time points that were further apart (Uher, 2011).

We excluded the switching task from further analysis for three main reasons. 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 Figure 9). Second, the correlations between the different measurement time points were low, suggesting no systematic individual differences (see Figure 10). 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 performance 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 replaced the switching task with the delay of gratification task in Phase 2.

**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 state factors based on the respective other test half. At time point 7, thresholds of both test-halves could not be set invariant across time (due to a divergent number of observed categories). Latent mean differences for the latent state variable at time point 7 should therefore be interpreted with caution.

Figure ?? [Manuel: das Figure label scheint nicht zu stimmen] visualizes the latent state means and reliability estimates from the LS model. Reliability was consistently high. None of the latent means was significantly different from zero, suggesting stable group level performance and no systematic mean change over time.

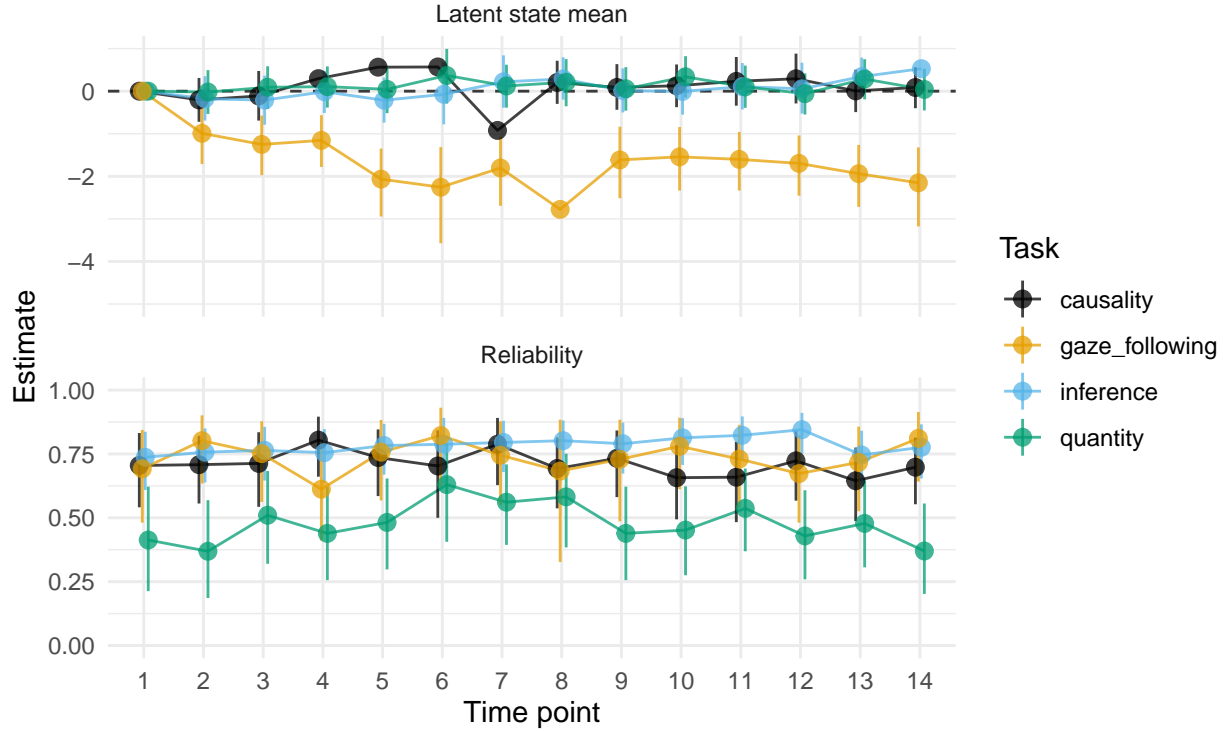


Figure 11: Latent means and reliability estimates with 95% CI for each time point based on LSM. Means at time point 1 are set to 0.

Figure ?? [Manuel: hier auch nicht] gives the correlations between the latent states for the different time points. Correlations were generally high, indicating stable individual differences.

In the LSTM, the consistency coefficient was estimated to be around .903. This means that around 90% of true inter-individual differences are attributable to stable (trait) differences between individuals, while approximately 10% are due to variance in time point specific deviations from the stable trait. Reliability (across time points) was estimated to be high with an mean of .725 (see Figure 13).

Figure 14 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 true inter-individual 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 variation, that is, variance in the time specific deviation of 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. As noted above – and as can be seen in Figure 9 – performance on a group level was clearly 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.

Reliability was high in the LS model and none of the latent means differed from zero (Figure ??). Correlations between latent states were 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 (trait) differences between individuals. Approximately 14% were due to variance

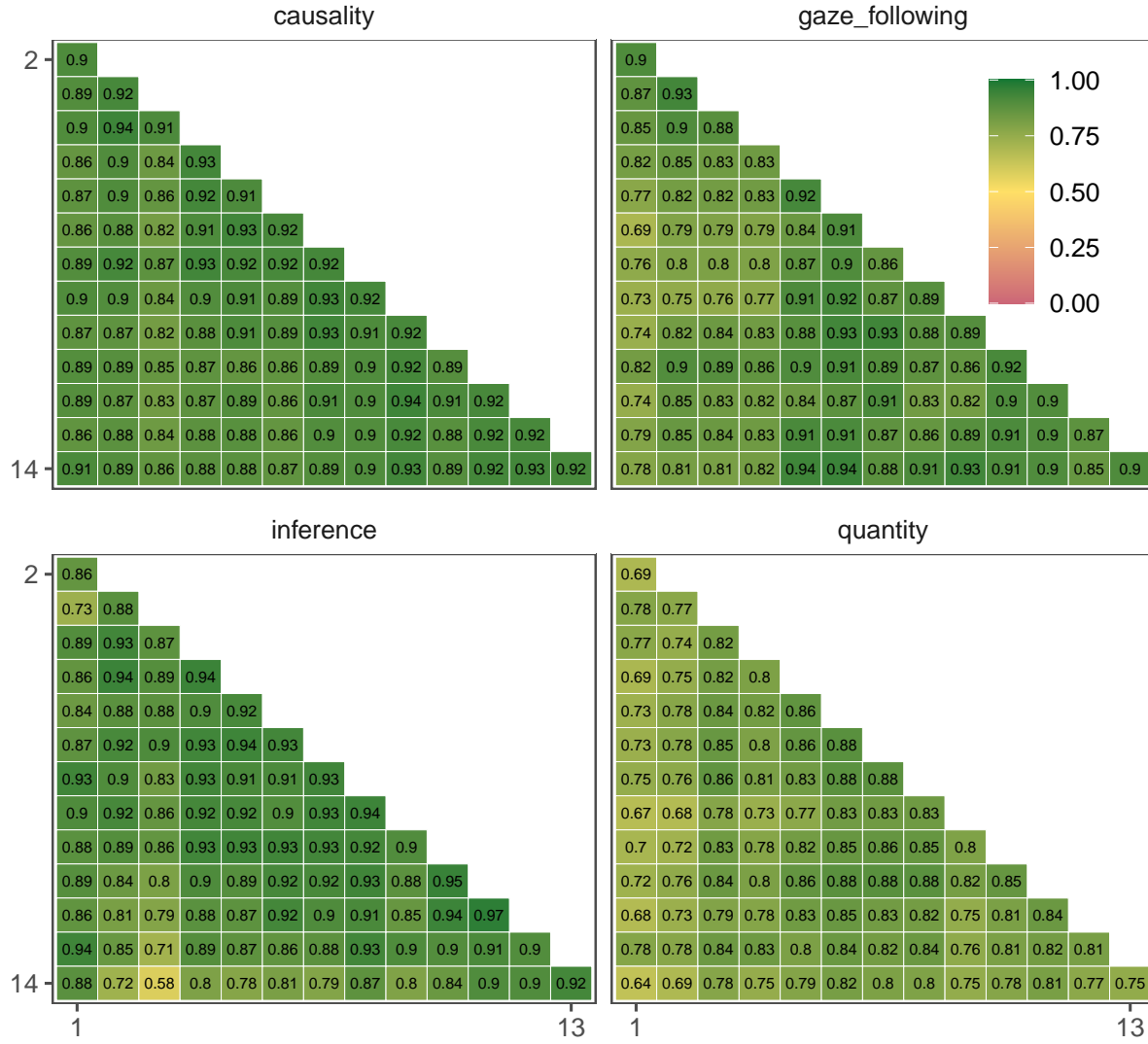


Figure 12: Correlations between latent state variables based on LSM for the different tasks.

in time-point specific deviations from the stable trait. Reliability was high with an estimate of .815 (see Figure 13).

According to the LSTM-AR, around 79.4% of true inter-individual differences at time point 14 went back to stable trait differences and around 8.3% of true inter-individual 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 saw a similar pattern as for the causal inference task: Performance was very stable on a group level and so were the differences between individuals. Interestingly, from Figure 9 we take that group-level performance was at chance. The stable individual differences we found here suggest that variation around this mean was systematic and therefore that some individuals consistently performed above chance. Thus, despite the fact that this task was very difficult for apes, it was suitable to measure individual differences.

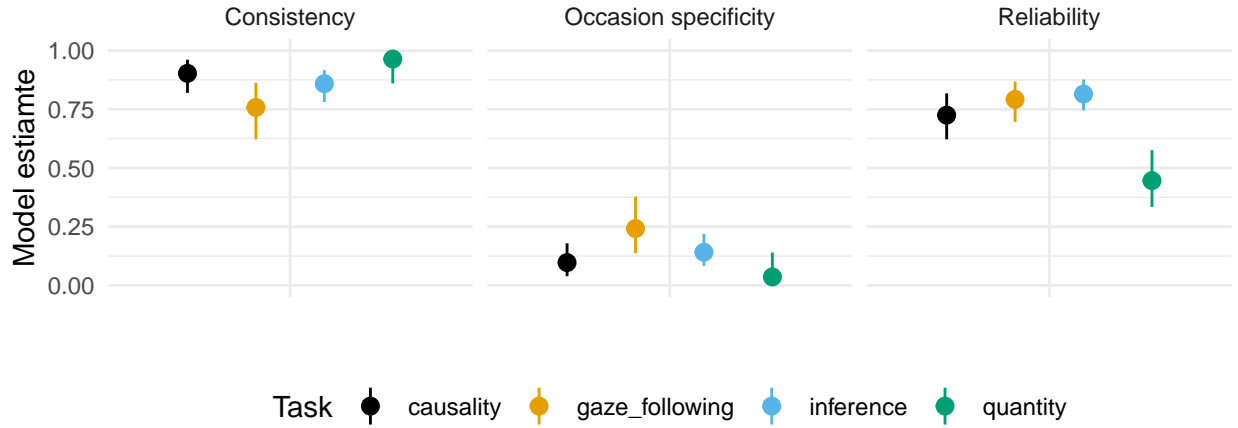


Figure 13: Model parameters (with 95% CI) from LSTM for the four tasks.

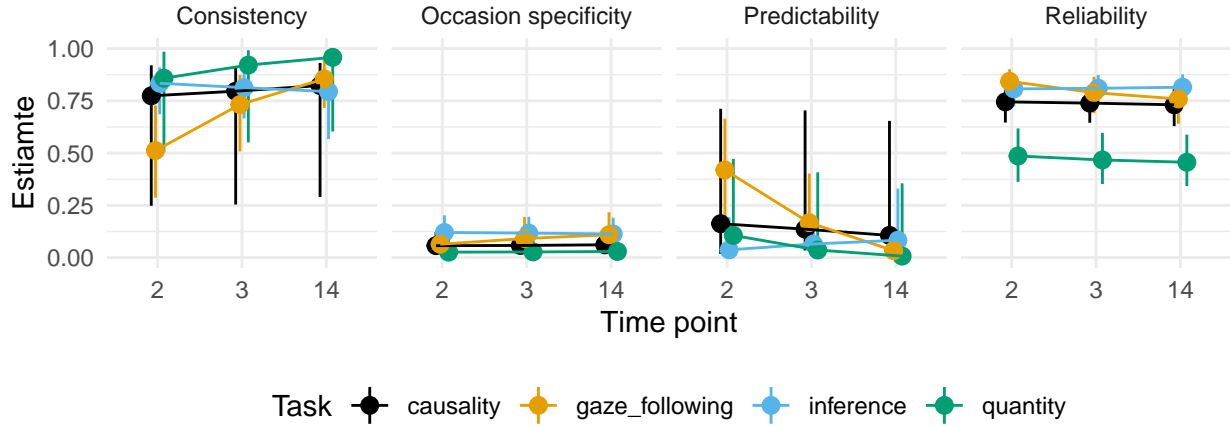


Figure 14: Model parameters (with 95% CI) from LSTM-AR for the four tasks.

**Gaze following** For gaze following, we had only 8 observed trials per measurement occasion. The highest two categories (3 and 4 correctly solved trials) were collapsed into one category due to sparsity. Thresholds could not be set equal for test half 2 at time point 9 as well as test half 1 and test half 2 at time point 8, 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 with the exception of time point 8.



Latent state means estimated in the LSM varied between -0.990 and -2.153 (for time point 8 the latent state mean was -2.77, but, as mentioned above, thresholds for this time point were not invariant). All of the latent state means were significantly lower than zero, suggesting a decrease in gaze following after the second time point (Figure 11). Reliability was high for all time points. The correlations between latent states for the different time points were generally high, pointing to stable individual differences (Figure 12).

In the LSTM, consistency was estimated to be around .758, that is around 76% of true inter-individual differences are attributable to stable differences between individuals. Approximately 24% of inter-individual differences were due to variance in time point specific deviations from the stable trait. Reliability was high with an estimate of .792.

According to the LSTM-AR, around 85.5% of true inter-individual differences at time point 14 went back to stable trait differences and around 3.3% of true inter-individual differences can be explained by carry-over effects from previous time points. Around 10.9% of the variance was due to time-specific variance between individuals. However, the state residual variance at time point 1 was estimated with great uncertainty (very large CI). Together with the very low predictability estimate, this suggests that this model is not particularly well suited to describe the results. [Manuel: zum letzten Satz: Ich bin mir nicht sicher, ob wir diese Interpretation so stehen lassen sollten. Das Modell passt laut PPP gut auf die Daten, der AR Effekt ist auch signifikant. Eventuell reicht eben die Stichprobengröße nicht, um die State Residual Varianz zu  $t=1$  mit größerer Genauigkeit zu schätzen. Was im Ergebnis auffällt ist die vergleichsweise große Varianz des State Residuums zu  $t=1$ , was anzeigt, dass dieser Zeitpunkt von den anderen verschieden zu sein scheint. Das spiegelt sich auch in den Korrelationen im LS Modell etwas wieder, man sieht in Figure 12 ganz gut, dass diese über die Zeit zunehmen. Eventuell ist das eine Art Gewöhnungseffekt? Bei  $t=1$  schauen die Affen noch häufiger hinterher, danach haben sie verstanden, dass es meist nichts zu sehen gibt, die Anzahl der gaze follows nimmt im Mittel ab und Unterschiede zwischen den Individuen werden geringer? wenn man sich die Rohwerte anschaut, sieht man dass pro Testhälfte ca. 2/3 der Affen ab  $t=5$  kein einziges mal mehr mit dem Blick folgen; wenn es immer die gleichen sind die es weiterhin tun / nicht mehr tun erklärt sich auch die Stabilität.]

In sum, we see a change in gaze following over time (Figure 9). This group-level effect, however, did not affect differences between individuals, which were systematic across time points.

**Quantity** The lowest three (out of seven possible) categories (0, 1 and 2 correctly solved trials per test half [Manuel: out of 6 ist richtig oder? ]) were collapsed due to sparsity. Thresholds could not be set equal for test half 1 at time point 5, 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.

Latent state means estimated in the LSM varied very little and all lay between -0.058 and 0.369. None of these state means differed from zero (Figure 11). Reliability estimates were substantially lower compared to the other tasks.

The consistency coefficient was estimated to be around .964, that is around 96% of true inter-individual differences was attributable to stable differences between individuals and only approximately 3.6% were due to variance in time-point specific deviations from the stable trait. Again, reliability was rather low with an estimate of .446.

According to the LSTM-AR, around 95.8% of true inter-individual differences at time point 14 went back to stable trait differences and only 0.7% of the inter-individual differences can be explained by carry-over effects from previous time points. The remaining 2.9% of the variance was due to time-specific variance between individuals.

Taken together, quantity judgments were very stable over time on the group level (see also Figure 9). The low reliability estimates suggest, however, that the task is less suited to capture individual differences.

## Phase 2

### Comparison between phases

### Relations between tasks

To analyse relations between different tasks (constructs), we estimated six separate LST models, each modeling the relation between two tasks. In these combined models, the sub-models for each task were equivalent to the LST models described above. For ease of model specification, the LST models were estimated as multilevel models. These models are equivalent to the LST models for single tasks under the assumption of strict factorial invariance. Figure 15 visualizes the model for two tasks.

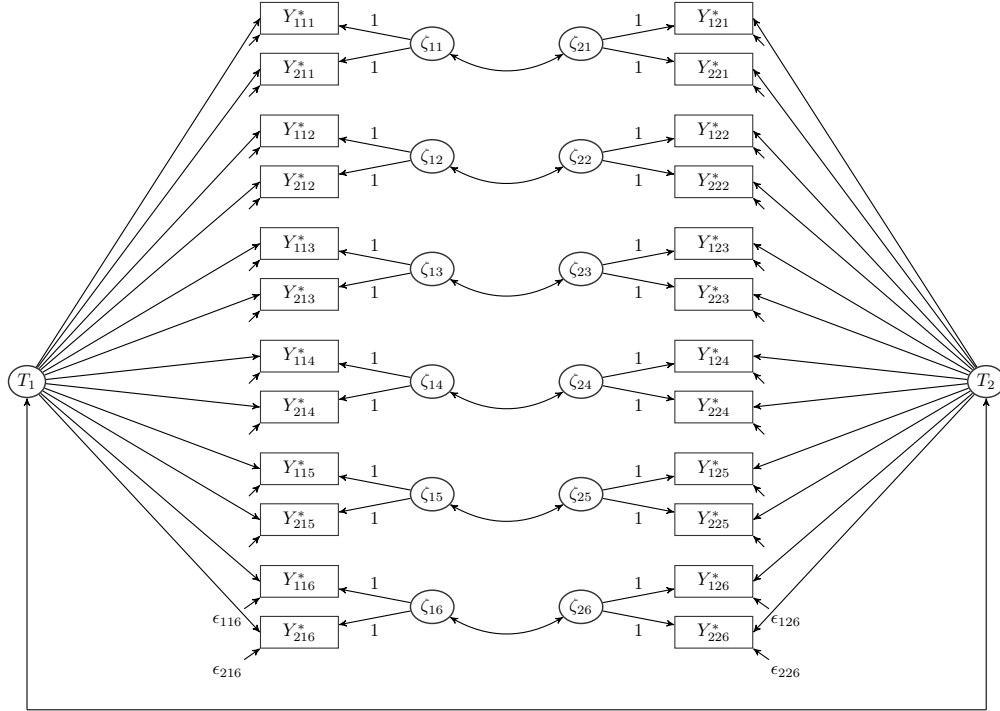


Figure 15: Latent State Trait models for two tasks with correlations between traits and states for two indicators and six measurement time points.

Detailed information on the parameter estimates obtained in LST models for each separate task is provided above. Here we report the results with a focus on the latent correlations only. The parameters of interest were correlations between a) the latent traits, indicating associations between stable cognitive ability as estimated by the different tasks, and b) correlations between state residual variables belonging to the same measurement time point, as an indicator of time-specific associations between latent abilities across the two tasks, above and beyond stable trait differences.

Simulation studies suggested that LST models in which latent correlations between time-specific state residual variables were estimated to be time-point specific (i.e. covariances and variances of state residual variables can freely vary across time) did not show good estimation performance under the given conditions (sample size, ordinal indicator variables, etc.). Therefore, we chose a model with fixed correlations between state residual variables across time. That is, a model in which we assumed that associations between latent time-specific cognitive abilities across two different tasks within each time point are equal at all time points. We think that this assumption is reasonable in the present context. As a consequence, just one correlation between latent states is estimated for each model. The corresponding model showed good estimation performance under the given sample sizes in a simulation study.

For details on MCMC estimation see section on estimation above. Because the multi-construct models were considerably more complex (i.e. had more parameters), we increased the minimum number of iterations per Markov chain to 20,000 (with a thinning of 10, that is, 200,000 iterations per chain).

## Phase 1

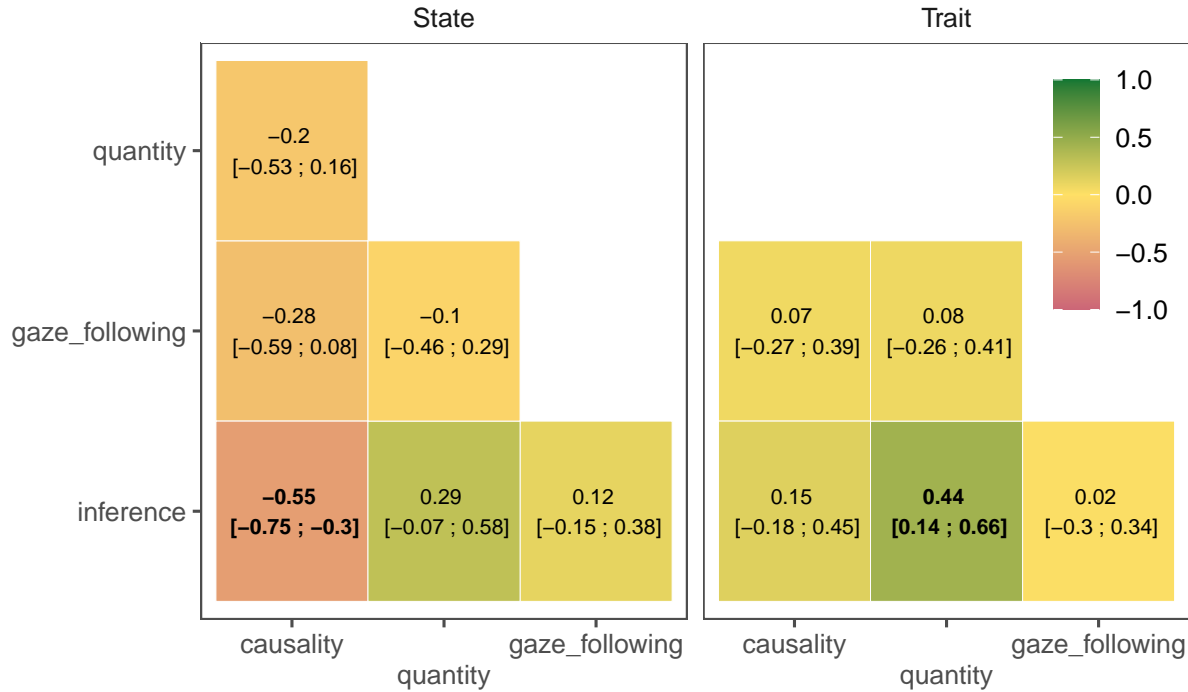


Figure 16: Correlations between latent traits and latent state residual variables, respectively, of different tasks. Bold correlations are different from zero as judged by the 95% CI. There is no trait correlation between quantity and causality because the corresponding model showed a poor fit. See main text for details.

[Manuel: bzgl. der figure note: die Korrelation zwischen quantity und causality wurde ja aber geschätzt, und das genutzte Modell zeigte guten Fit (wie in der Tabelle zu sehen). Vielleicht umformulieren zu: “The trait correlations between quantity and causality are not displayed, see main text for details ?” bzgl. der Figure selbst: könntest du aus “State” in der Überschrift bitte “State Residual” machen? States sind die aus dem LS Modell...]

Model fit indices are shown in Table 3. Due to a low PPP value, the model for causality and quantity was modified such that for each task, test-half specific trait factors were estimated on the between-level. The correlations between the two tasks are therefore also reported as test-half specific trait correlations.

The only correlations for which the 95% CI did not include zero were the state residual correlation between causality and inference ( $r_{sc,si} = -0.551$ , 95% CI = [-0.749; -0.299]) and the trait correlation between inference and quantity ( $r_{ti,tq} = 0.436$ , 95% CI = [0.135; 0.665]).

The negative state correlations between causality and inference may be explained by the way the two tasks were presented. Remember that causality and inference trials used the same setup and were intermixed. A negative correlation suggests that higher (residual) performance in one task was associated with lower performance in the other task. Responding correctly in the two tasks required opposite choice behaviors. That is, in causality, the ape had to pick the cup the experimenter shook to be correct. In inference, it was the unshaken cup. Such a negative correlation arises when sometimes participants respond in the same way (e.g. pick the shaken cup) across tasks. Note, however, that if this were a stable strategy which individuals would consistently use, we would have seen a negative correlation between the trait estimates. The best

Table 3: Model fit indices for multi-construct models

Task1	Task2	PPP	Chi 95% CI
causality	gaze following	0.371	-17.73 ; 24.37
	inference	0.273	-14.83 ; 28.64
	quantitiy	0.419	-19.26 ; 23.54
inference	gaze following	0.419	-18.55 ; 24.28
	quantitiy	0.341	-16.10 ; 25.86
quantitiy	gaze following	0.402	-18.96 ; 23.43

*Note:*

PPP = Posterior predictive p-value

Chi 95% CI = 95%CI of difference between predicted and observed chi-square values

explanation is thus that there are short periods of inattentiveness during which (some) participants confused the two tasks.

The trait correlation between inference and quantity was positive, suggesting that individuals with better quantity judgment abilities also have better inferential abilities.

One (out of four) of the test-half specific trait correlations between causality and quantity was also reliably different from zero ( $r_{tc2,tq1} = 0.436$ , 95% CI = [0.135; 0.665]). We do not consider this result to be substantial evidence for a substantive association between the trait estimates in the two tasks and therefore do not interpret it any further. Figure 16 shows all correlations between the different tasks.

## Phase 2

### Comparison between phases

### Predictability

The output of the projection predictive inference models is a ranking of the different predictors with respect to how much they improve a model’s fit. Predictors ranked first improve the fit the most, while later predictors yield smaller improvements (if any). The selection of “relevant” predictors is based on plotting the loss statistics and looking for a point at which it levels off. As such, the selection is to some extent arbitrary. The ranking, however, is not. When we compare the results from the two phases, we not just look at which subset of predictors is selected, but also at the overall ranking.

## Phase 1

**Causal inference** Figure 17 visualizes the results. Out of the 13 predictor variables we analysed, we selected only **group** to be relevant in addition to the random intercept term. When inspecting the projected posterior distribution for **group**, we saw substantial differences between the groups: Orangutans and the b-chimpanzee group performed best, followed by bonobos and finally the gorillas and the a-chimpanzee group (see Figure 17B).

**Inference by exclusion** Figure 18 visualizes the results. For inference by exclusion, we selected **time\_in\_leipzig**, **group**, and **age** as relevant predictors in addition to the random intercept term. All three predictors capture stable individual characteristics.

Figure 18B shows the projected posterior distributions for the predictors and suggests that the longer apes lived in Leipzig, the better their performance was. The differences between groups were such that the two

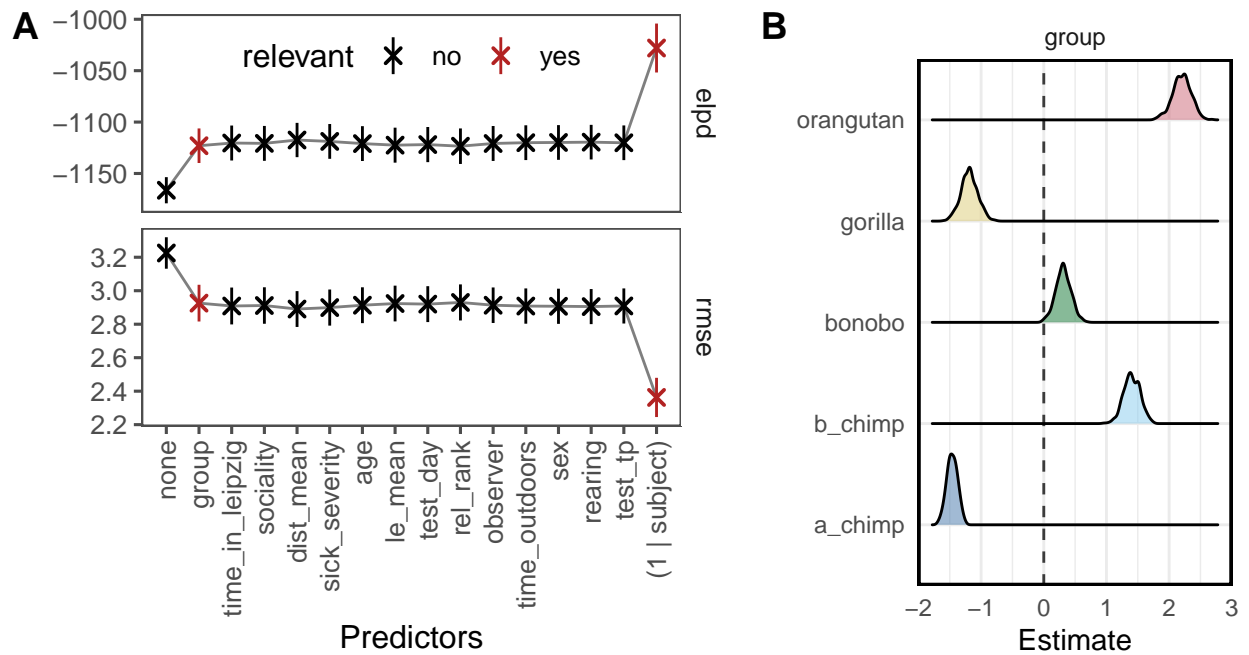


Figure 17: Predictor selection for causality. A) Elpd and RMSE values for predictors, ordered by importance (left to right) according to the cross-validated projection prediction model. Note that the random intercept term was forced to be the last one to be included. B) Projections for the selected predictors based on the submodel.

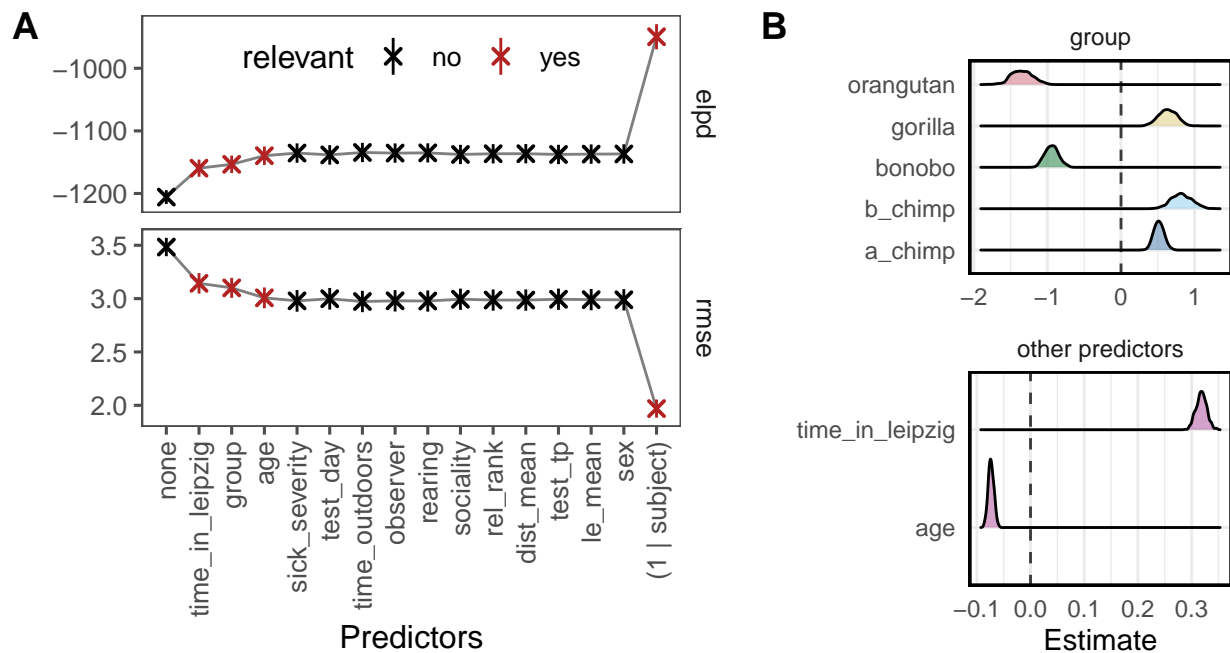


Figure 18: Predictor selection for inference. A) Elpd and RMSE values for predictors, ordered by importance (left to right) according to the cross-validated projection prediction model. Note that the random intercept term was forced to be the last one to be included. B) Projections for the selected predictors based on the submodel.

chimpanzee groups together with the gorillas performed on a higher level compared to the bonobos and orangutans. With respect to **age**, we found that performance decreased with age.

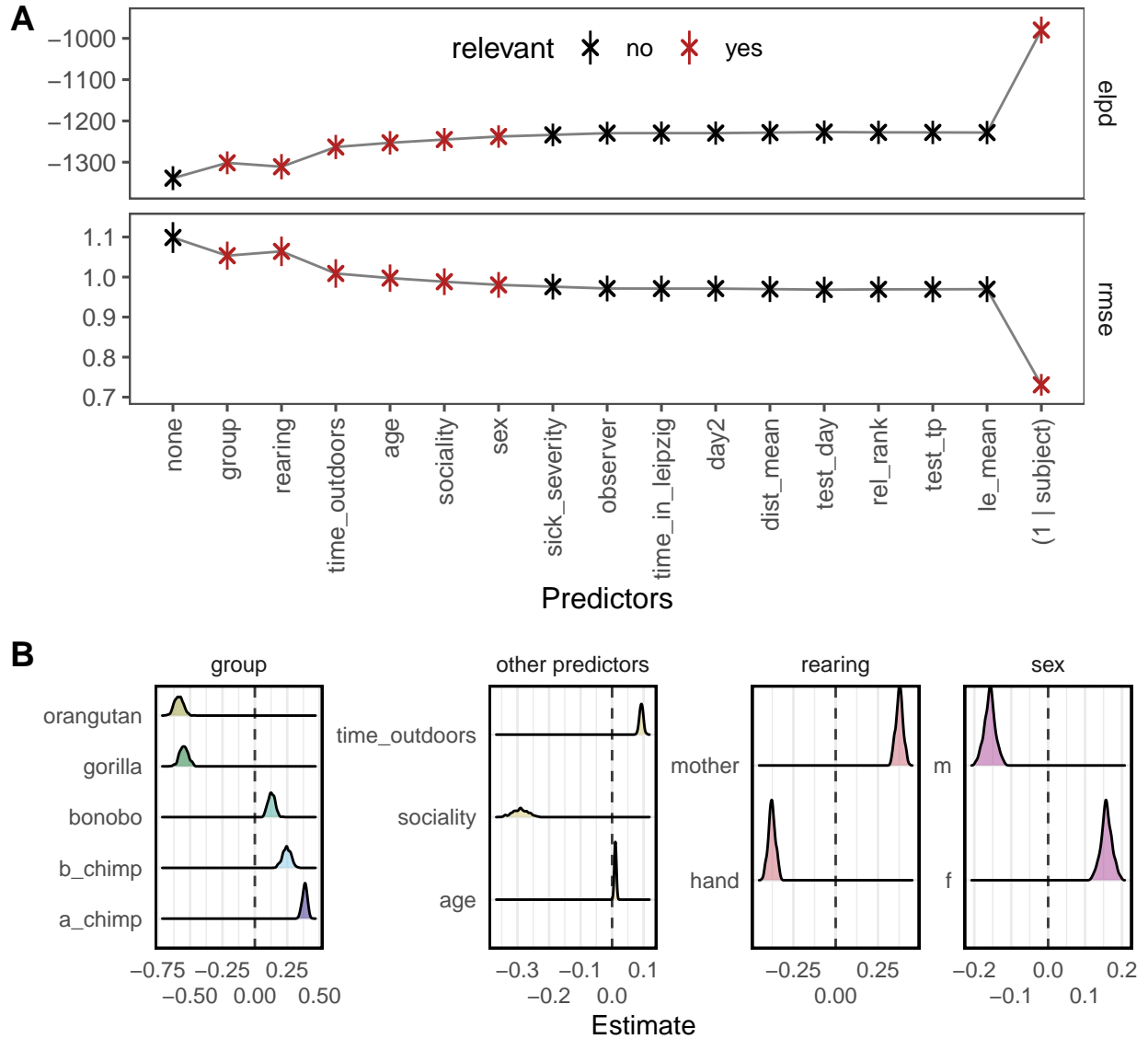


Figure 19: Predictor selection for gaze following. A) Elpd and RMSE values for predictors, ordered by importance (left to right) according to the cross-validated projection prediction model. Note that the random intercept term was forced to be the last one to be included. B) Projections for the selected predictors based on the submodel model.

**Gaze Following** Figure 19 visualizes the results. Gaze following had the most selected predictors of all tasks. In addition to the random intercept term, we selected, **group**, **rearing**, **time\_outdoors**, **age**, **sociality**, and **sex**. Again, most of these predictors were stable individual characteristics, with the exception of **time\_outdoors** and **sociality**.

Groups differed in that a-chimpanzees were most likely to follow gaze, followed by b-chimpanzees and bonobos. Gorillas and orangutans were the least likely to follow the experimenter's gaze. Mother-reared individuals outperformed hand-reared individuals (including those with an unknown rearing history). The more time individuals spent outdoors, the more likely they were to follow gaze. Also, the probability to follow

gaze increased with age. Individuals with a lower sociality index had higher rates of gaze following. Finally, females outperformed males. Figure 19B visualizes these results.

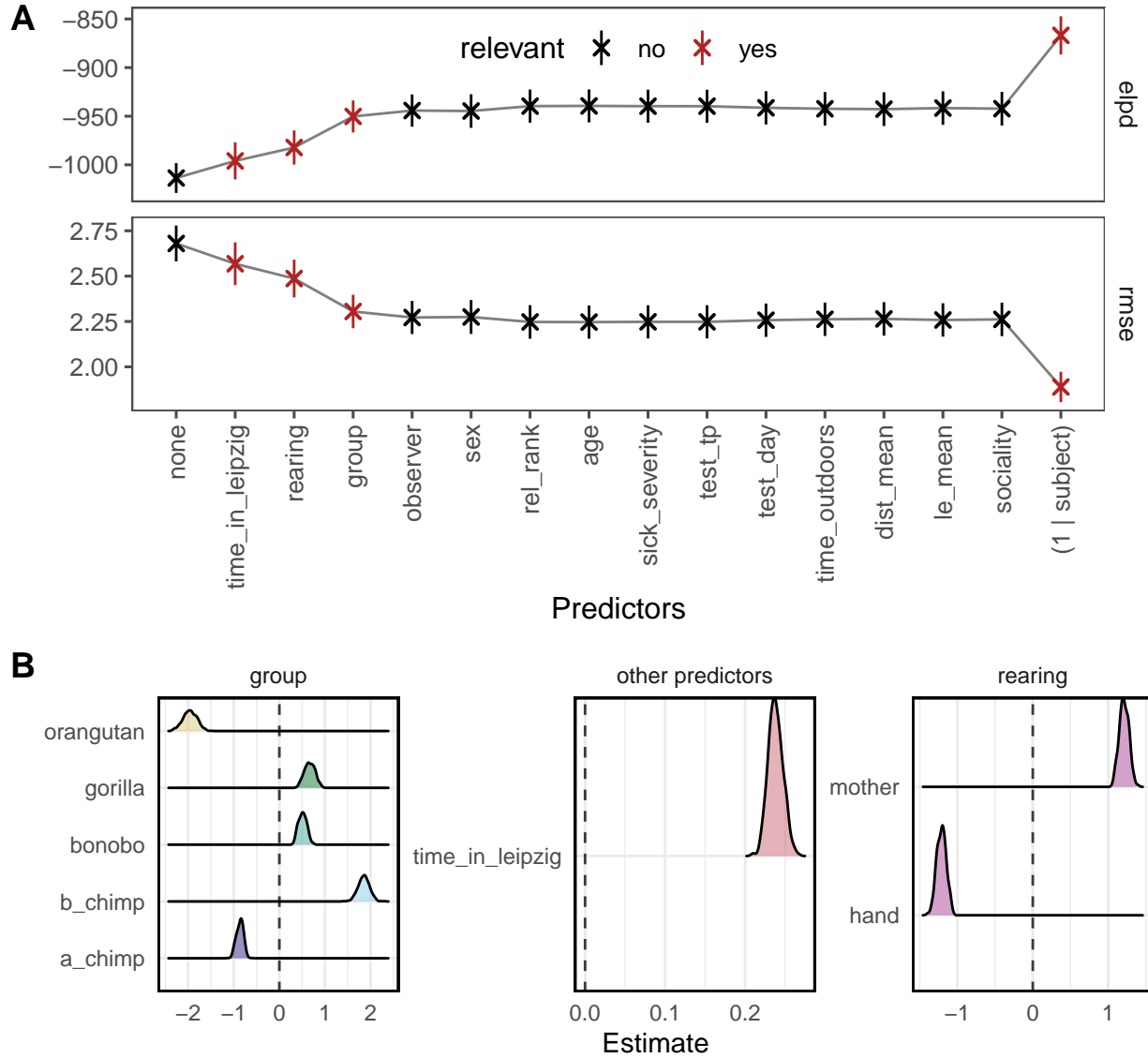


Figure 20: Predictor selection for quantity. A) Elpd and RMSE values for predictors, ordered by importance (left to right) according to the cross-validated projection prediction model. Note that the random intercept term was forced to be the last one to be included. B) Projections for the selected predictors based on the submodel model.

**Quantity** Figure 20 visualizes the results. For quantity, we selected three predictors in addition to the random intercept term: `time_in_leipzig`, `rearing`, and `group`. All of these predictors were stable individual characteristics.

The longer individuals had lived in Leipzig, the better they performed in the task. Group differences were such that b-chimpanzees performed best, followed by bonobos and gorillas. A-Chimpanzees performed slightly worse, but still better than the orangutans. Once again, mother reared individuals outperformed those who were hand-reared or whose rearing history was unknown. Figure 20B visualizes these results.

**Summary** The most obvious result was that the random intercept term ( $1 \mid \text{subject}$ ) was – by far – the predictor that improved the model fit the most. This suggests that a large portion of the variance is explained by stable individual characteristics that we did not capture in our predictors. Most likely, these are the outcomes of idiosyncratic developmental processes or genetic pre-dispositions, which operate on a much longer time-scale than what we captured in our study.

Second, we saw that most of the relevant predictors came from the group of stable individual characteristics. This aligns well 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. Following this reasoning, there was very little *systematic* variation between time points, and thus not much the time-varying predictors could account for. In line with this interpretation, we selected time point specific predictors only for gaze following, the task with the highest occasion specificity estimate according to the LSTM.

The predictor selected most often was **group**. Differences between groups were, however, variable. The b-chimpanzee group tended to perform best across tasks, but the ranking of the other groups (including the other chimpanzee group) changed from task to task. This speaks against clear species differences in general cognitive performance. Again, the most likely explanation for group differences is an interaction between species specific dispositions and individual- / group-level developmental processes.

The predictors that were selected more than once influenced performance in a systematic way. Whenever rearing history was selected to be relevant, mother-reared individuals outperformed others. The more time an individual had lived in Leipzig, the better performance was. An exception was **age**, which had a positive estimate for gaze following but a negative one for inference.

When zooming out, we found no clear ranking of predictors across tasks (see Figure 21). It is important to note, however, that for higher ranks, the difference between ranks in the loss statistic is very small and the ordering to some extent arbitrary. But even if we focus only on the five highest-ranked predictors per task, we see a lot of variation across tasks – with **group** being a notable exception.

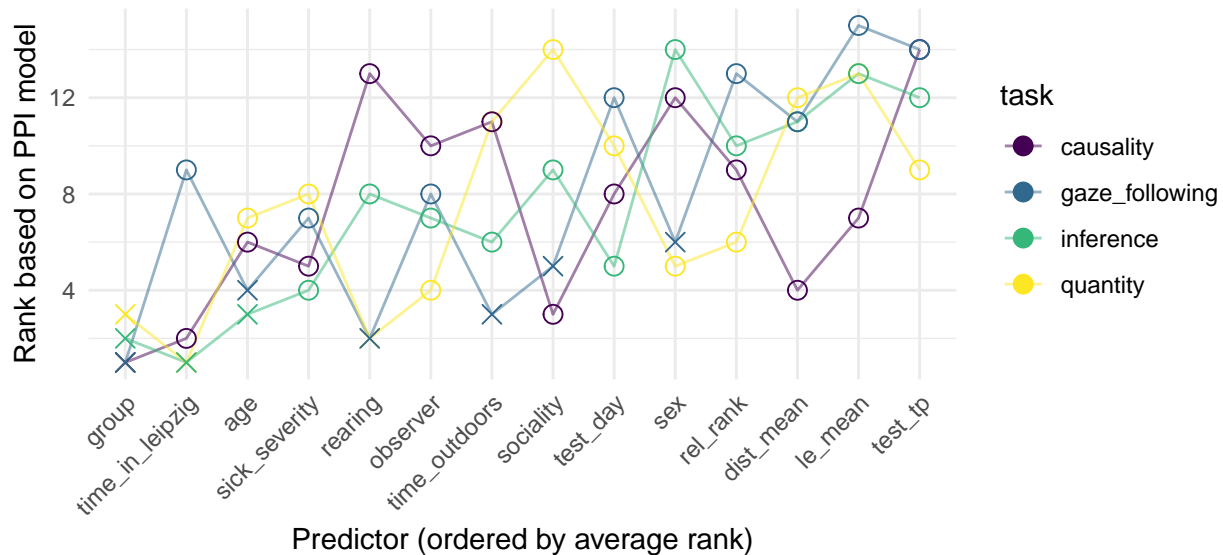


Figure 21: Ranking for each predictor and task. Crosses denote selected predictors.



## Phase 2

### Comparison between phases

## Summary

## References

- Bollen, K. A. (1989). *Structural equations with latent variables* (Vol. 210). John Wiley & Sons.
- 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.
- Geiser, C. (2020). *Longitudinal structural equation modeling with mplus: A latent state-trait perspective*. Guilford Publications.
- 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.
- 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>
- Leckie, G. (2019). Multiple membership multilevel models. Retrieved from <https://arxiv.org/abs/1907.04148>
- Pavone, F., Piironen, J., Bürkner, P.-C., & Vehtari, A. (2020). Using reference models in variable selection. Retrieved from <https://arxiv.org/abs/2004.13118>
- 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>
- 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.
- Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika*.
- 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.
- 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*.
- 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.
- 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.
- 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.

# Appendix

## SEM Simulations

### Simulation setup

Data were generated and estimated using MPlus 8.4. Data-generating values are based on the real-data application of the models to the available subset of the data at the time of conducting the simulation study. That is, data were simulated for 40 individuals (N) observed across 9 or 12 measurement occasions, with 5 or 7 observed categories per indicator. 1000 replications were simulated. Data estimation took place using the MPlus default priors. In case of LST models for one construct, default priors were compared with IG(0.001, 0.001) priors set on all variance parameters (model did not include latent covariances). Two chains with a minimum of 5,000 iterations per chain and a thinning factor of 10 was applied (i.e. at last 50,000 iterations of which only every 10th was used for constructing the posterior distribution). Convergence was assumed and estimation stopped when the PSR fell below 1.05 for the first time after the minimum number of iterations was reached.

### Simulation results

In the following, the 95% coverage rate, the Relative Parameter Estimation Bias (deviation between average estimate and population parameter divided by the population parameter), the Mean Squared Error, absolute bias, as well as Relative Standard Error Bias are displayed for every simulated model (Figure 22 - 30). Relative parameter and standard error biases below 0.1 (that is  $< 10\%$ ) are considered acceptable.

Parameters in the latent state models for one construct are estimated accurately, with relative biases below a cutoff of 10% bias and good coverage rates, irrespective of simulating 7 or 5 observed ordered categories for the observed indicators.

Latent State Trait models for one construct with latent state residual variances fixed across time show good estimation performance, with both default or adapted inverse gamma priors. When freely estimating latent state residual variances across time points (i.e., no restrictions on variances), model parameters are not estimated accurately under the simulated sample sizes, irrespective of the prior choice.

Latent State Trait models for a combination of two constructs with latent variances and covariances of the state residual variances restricted to equality across time points work well. Models with freely estimated variances and covariances do not show good estimation performance. The same holds for the Latent State model with two constructs and freely estimated variances and covariances.

In conclusion, Latent State models for one construct (freely estimated variances) as well as Latent State Trait models for one or two constructs with state residual variances restricted across time exhibit good estimation performance (low biases, high coverage) and application under the simulated samples size should be feasible in practice.

## Latent State models: One construct

Freely varying state variances and covariances across time points. 5 ordered categories

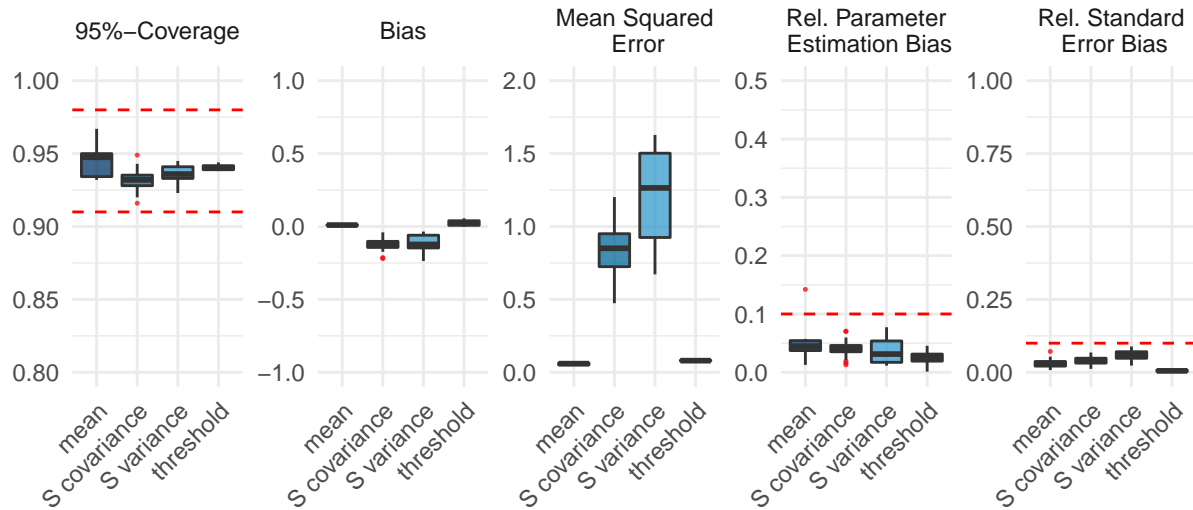


Figure 22: Results of the simulation study for the Latent State (LS) model including one construct with freely estimated latent State variances and covariances, spanning 9 measurement time points. Ordinal indicators were simulated with 5 ordered categories. Boxplots display the distribution of the respective statistic across different parameters of the same parameter type.

## Latent State models: One construct

Freely varying state variances and covariances across time points. 7 ordered categories

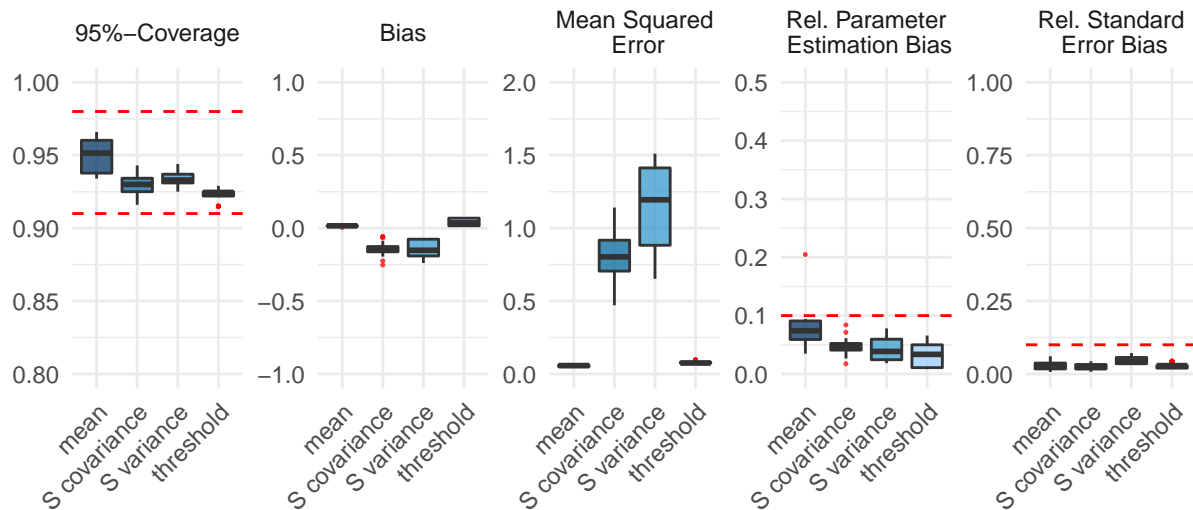


Figure 23: Results of the simulation study for the Latent State (LS) model including one construct with freely estimated latent State variances and covariances, spanning 9 measurement time points. Ordinal indicators were simulated with 7 ordered categories. Boxplots display the distribution of the respective statistic across different parameters of the same parameter type.

## Latent State Trait models: One construct

Fixed state residual variances across time points with default priors

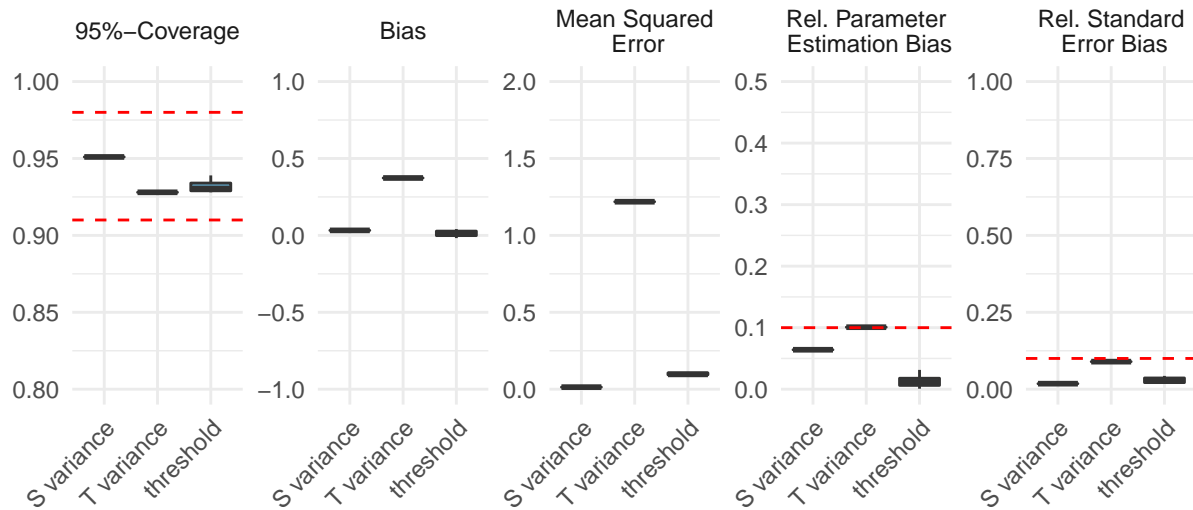


Figure 24: Results of the simulation study for the Latent State Trait (LST) model including one construct with latent state residual variances fixed to be equal across time, spanning 9 measurement time points. MPlus default priors. Ordinal indicators were simulated with 7 ordered categories. Boxplots display the distribution of the respective statistic across different parameters of the same parameter type.

## Latent State Trait models: One construct

Fixed state residual variances across time points with inverse gamma priors

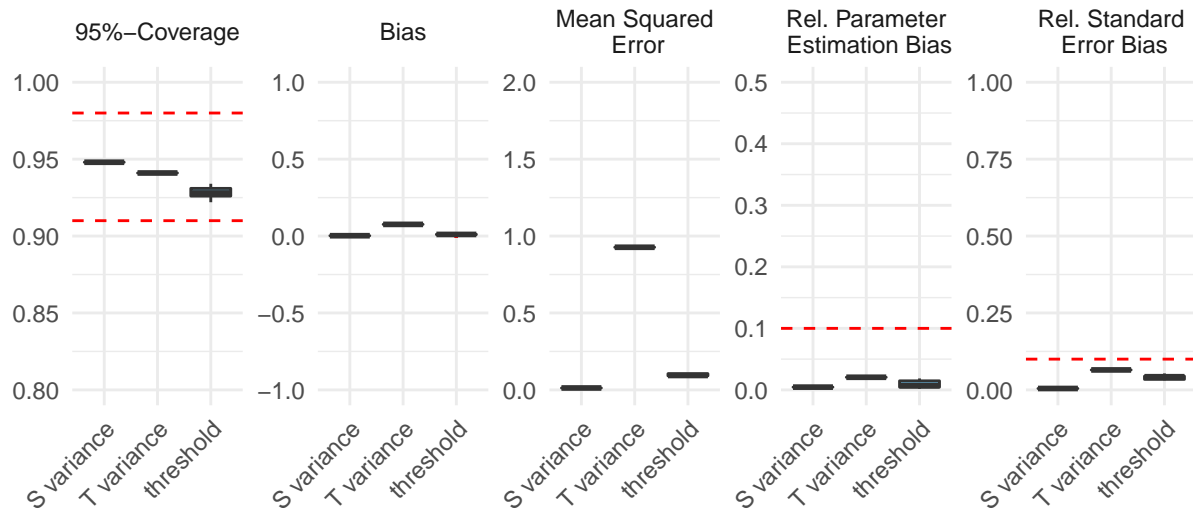


Figure 25: Results of the simulation study for the Latent State Trait (LST) model including one construct with latent state residual variances fixed to be equal across time, spanning 9 measurement time points. Inverse gamma priors  $IG(0.001, 0.001)$  for all variances. Ordinal indicators were simulated with 7 ordered categories. Boxplots display the distribution of the respective statistic across different parameters of the same parameter type.

## Latent State Trait models: One construct

Free state residual variances across time points with default priors

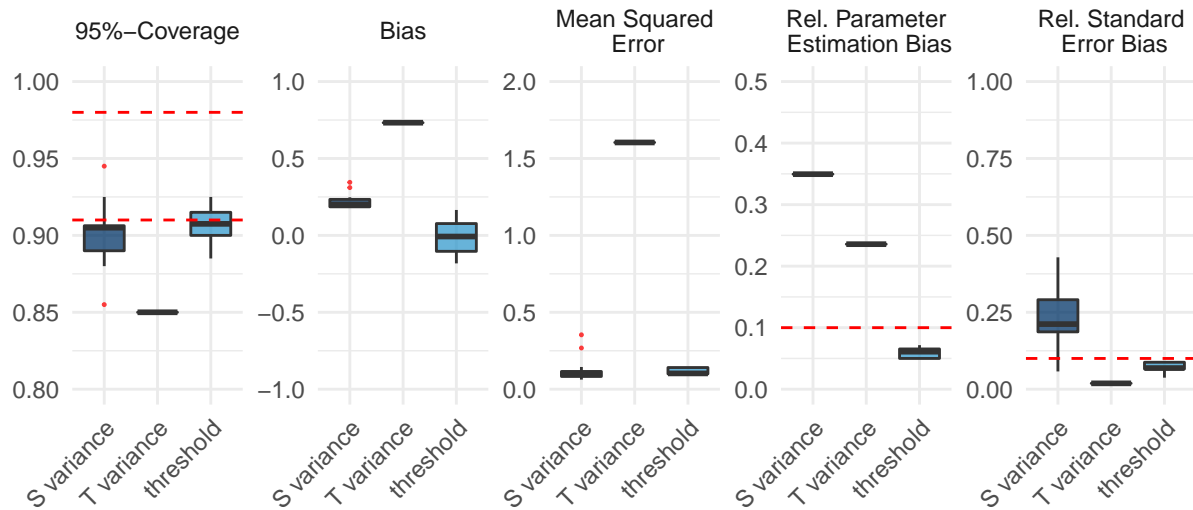


Figure 26: Results of the simulation study for the Latent State Trait (LST) model including one construct with latent state residual variances freely estimates across time, spanning 9 measurement time points. Ordinal indicators were simulated with 7 ordered categories. Boxplots display the distribution of the respective statistic across different parameters of the same parameter type.

## Latent State Trait models: One construct

Free state residual variances across time points with inverse gamma priors

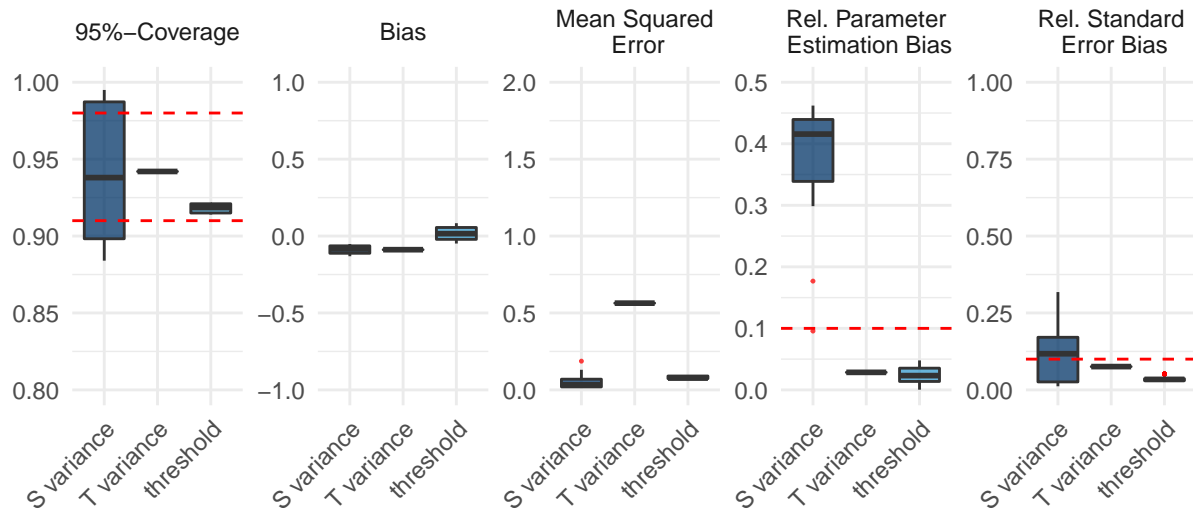


Figure 27: Results of the simulation study for the Latent State Trait (LST) model including one construct with latent state residual variances freely estimates across time, spanning 9 measurement time points. Inverse gamma priors  $IG(0.001, 0.001)$  for all variances. Ordinal indicators were simulated with 7 ordered categories. Boxplots display the distribution of the respective statistic across different parameters of the same parameter type.

## Latent State models: Two constructs

Free state variances and covariances across time points and default priors

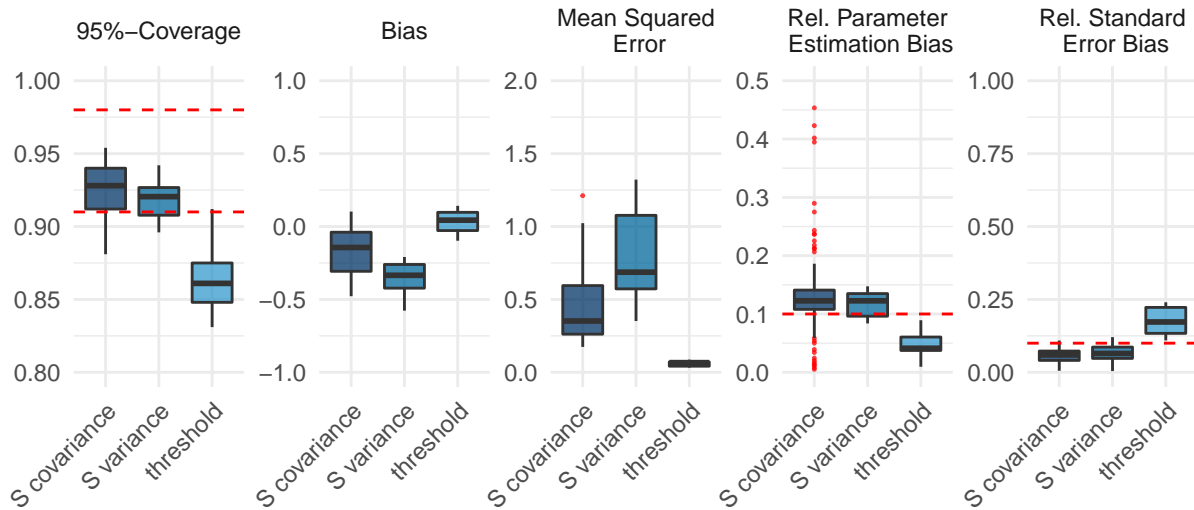


Figure 28: Results of the simulation study for the Latent State model including two constructs with latent state variances freely estimates across time, spanning 9 measurement time points. Ordinal indicators were simulated with 7 ordered categories. Boxplots display the distribution of the respective statistic across different parameters of the same parameter type.

## Latent State Trait models: Two constructs

Free state residual variances and covariances across time points

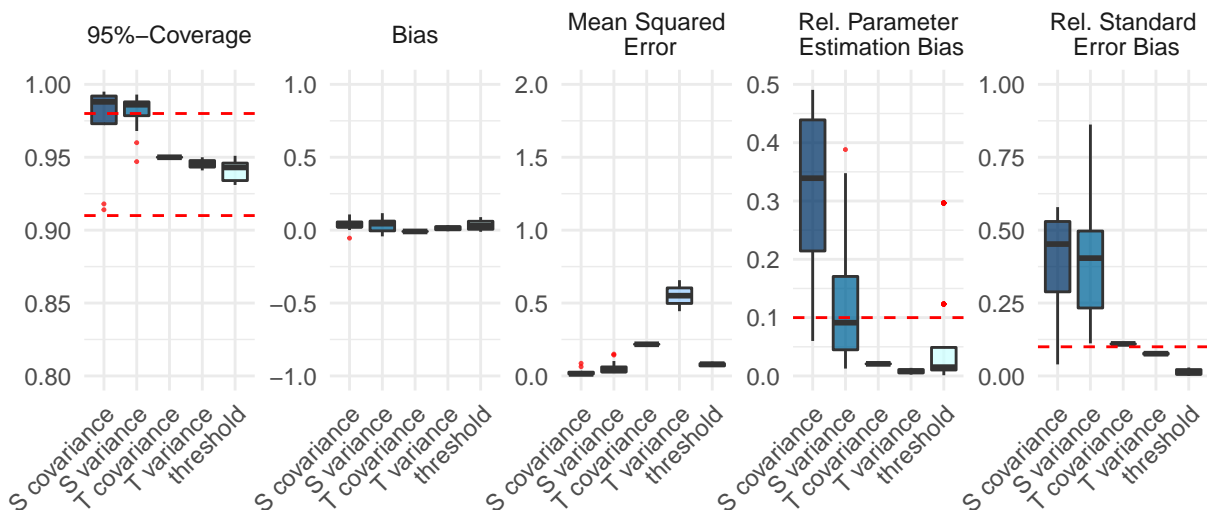


Figure 29: Results of the simulation study for the Latent State Trait (LST) model including two constructs with free latent state residual variances across time, spanning 9 measurement time points. Ordinal indicators were simulated with 7 ordered categories. Boxplots display the distribution of the respective statistic across different parameters of the same parameter type.

## Latent State Trait models: Two constructs

Fixed state residual variances and covariances across time points

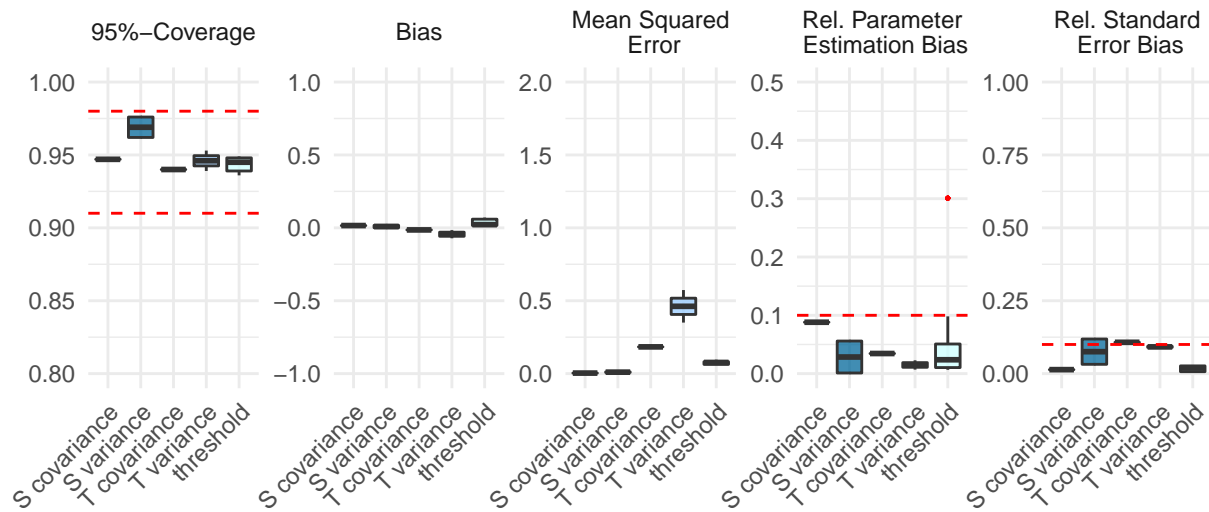


Figure 30: Results of the simulation study for the Latent State Trait (LST) model including two constructs with fixed latent state residual variances across time, spanning 9 measurement time points. Ordinal indicators were simulated with 7 ordered categories. Boxplots display the distribution of the respective statistic across different parameters of the same parameter type.