

¹ Individual differences in great ape cognition across time and domains: stability, structure,
² and predictability

³ Manuel Bohn^{1,2}, Christoph Völter², Daniel Hanus², Nico Eisbrenner², Johanna Eckert², Jana
⁴ Holtmann³, & Daniel Haun²

⁵ ¹ Institute of Psychology in Education, Leuphana University Lüneburg

⁶ ² Department of Comparative Cultural Psychology, Max Planck Institute for Evolutionary
⁷ Anthropology, Leipzig, Germany

⁸ ³ Wilhelm Wundt Institute of Psychology, Leipzig University, Leipzig, Germany

¹⁰ Manuel Bohn was supported by a Jacobs Foundation Research Fellowship
¹¹ (2022-1484-00). We are grateful to thank all children and caregivers for participating in the
¹² study. We thank the Max Planck Society for the Advancement of Science.

¹³ The authors made the following contributions. Manuel Bohn: Conceptualization,
¹⁴ Formal Analysis, Writing - Original Draft Preparation, Writing - Review & Editing;
¹⁵ Christoph Völter: Conceptualization, Writing - Original Draft Preparation, Writing - Review
¹⁶ & Editing; Daniel Hanus: Conceptualization, Writing - Original Draft Preparation, Writing -
¹⁷ Review & Editing; Nico Eisbrenner: Formal Analysis, Writing - Original Draft Preparation,
¹⁸ Writing - Review & Editing; Johanna Eckert: Conceptualization, Writing - Original Draft
¹⁹ Preparation, Writing - Review & Editing; Jana Holtmann: Formal Analysis, Writing -
²⁰ Original Draft Preparation, Writing - Review & Editing; Daniel Haun: Conceptualization,
²¹ Writing - Review & Editing.

²² Correspondence concerning this article should be addressed to Manuel Bohn,
²³ Universitätsallee 1, 21335 Lüneburg, Germany. E-mail: manuel.bohn@leuphana.de

24

Abstract

25 Variation in cognitive abilities is critical to understanding both the evolution and
26 development of cognition. In this study, we examined the stability, structure, and
27 predictability of individual differences in cognitive abilities in great apes across a broad
28 range of domains, including social cognition, reasoning about quantities, executive functions,
29 and inferential reasoning. We repeatedly administered six tasks to $N = 48$ apes from all four
30 great ape species, spanning 10 sessions over 1.5 years. Results revealed substantial variability
31 across tasks. Some tasks exhibited significant improvements in group-level performance over
32 time, while others remained stable. The precision with which tasks measured individual
33 differences also varied, ranging from very poor to very good. Stable differences in cognitive
34 abilities between individuals explained a large proportion of the variance, yet these
35 differences were weakly associated across tasks. Task performance was most strongly
36 predicted by stable, individual-specific characteristics rather than transient variables. Our
37 findings highlight the malleability of great ape cognition, its multidimensional nature, and
38 the need for longitudinal research to uncover the ontogenetic origins of individual differences.

39 *Keywords:* great apes, cognition, individual differences

40 Individual differences in great ape cognition across time and domains: stability, structure,
41 and predictability

42 **Introduction**

43 Variation fuels evolution. Individual differences in cognitive abilities are essential for
44 understanding what evolves (Shaw & Schmelz, 2017; Thornton & Lukas, 2012; Völter,
45 Tinklenberg, Call, & Seed, 2018). These differences reveal which aspects of cognition are
46 invariant and which are malleable. They also shed light on the broader structure of the
47 cognitive architecture by identifying relationships between different cognitive abilities.
48 Moreover, they help identify the socio-ecological factors shaping cognition during both
49 ontogeny and phylogeny.

50 Broadly speaking, great ape cognition is marked by substantial individual variability
51 across domains, such as tool use, communication, social cognition, causal reasoning, and
52 reasoning about quantities. This variability has been observed in both captive and wild
53 settings (Berdugo, Cohen, Davis, Matsuzawa, & Carvalho, 2023; Bohn et al., 2023; Fröhlich,
54 Wittig, & Pika, 2019; Herrmann, Hernández-Lloreda, Call, Hare, & Tomasello, 2010; Watson
55 et al., 2018). Such findings suggest significant plasticity in cognitive abilities, shaped by
56 social and ecological influences. As noted above, such individual differences can be used to
57 study the structure of great ape cognition and its origin (Völter et al., 2018).

58 Despite their importance, few studies have explored the broader structure of individual
59 differences in great apes. Most work has focused on finding something akin to general
60 intelligence or a *g*-factor (Burkart, Schubiger, & Schaik, 2017; Deaner, Van Schaik, &
61 Johnson, 2006; Reader, Hager, & Laland, 2011). Using the Primate Cognition Test Battery
62 (PCTB) (Herrmann, Call, Hernández-Lloreda, Hare, & Tomasello, 2007), Herrmann et al.
63 (2010) found no evidence for a single g-factor in chimpanzees. Instead, they observed a
64 bifactorial structure, with one factor linked to spatial tasks and the other to social and
65 physical tasks. Similar findings have been reported for other primates (Fichtel, Dinter, &

66 Kappeler, 2020; Schmitt, Pankau, & Fischer, 2012). By contrast, Hopkins, Russell, and
67 Schaeffer (2014) used the PCTB to test a different sample of chimpanzees and identified a
68 *g*-factor, which was later found to relate to measures of self-control (Beran & Hopkins, 2018).
69 However, this study did not test whether the proposed structure (a single *g*-factor) fit the
70 data well. In a subsequent re-analysis, Kaufman, Reynolds, and Kaufman (2019) combined
71 data sets collected with the PCTB and found the single *g*-factor model inadequate. Only
72 multidimensional models accurately described the data. Beyond general cognitive abilities,
73 Völter, Reindl, et al. (2022) investigated the structure of executive functions in chimpanzees
74 using a multi-trait, multi-method approach. Their results showed limited evidence for the
75 structure proposed for executive functions in humans.

76 The existence of individual differences raises questions about their origins. Most
77 theories about the factors influencing the emergence of complex cognitive abilities operate on
78 a species level (Dunbar & Shultz, 2017; Henke-von der Malsburg, Kappeler, & Fichtel, 2020;
79 Rosati, 2017). Empirical studies in this tradition often compare closely related species with
80 differing social structures or ecological pressures (Amici, Aureli, & Call, 2008; Joly et al.,
81 2017; Kaigaishi, Nakamichi, & Yamada, 2019; Rosati & Santos, 2017). Alternatively,
82 researchers aggregate data across studies to compare species on a larger scale (Deaner et al.,
83 2006; Piantadosi & Kidd, 2016). This approach, however faces challenges in comparability,
84 as data are often collected using inconsistent methods (Schubiger, Fichtel, & Burkart, 2020).
85 An exception is ManyPrimates et al. (2022), which employed standardized methods to
86 collect a large data set on short-term memory and test species-level hypotheses. However,
87 their results were sobering: no single socio-ecological predictor explained cognitive variation
88 better than phylogenetic relatedness.

89 In contrast, much less research has focused on the individual level (Sih, Sinn, &
90 Patricelli, 2019). Early work focused on the effects of enculturation – raising great apes in a
91 human environment. Most of these studies, however, involved only one individual, making it

92 difficult to identify the relevant aspects experience that led to the observed changes in
93 cognition (see Berio & Moore, 2023 for a recent summary). Few studies with larger samples
94 exist: Watson et al. (2018) found that hand-reared chimpanzees are more likely to use social
95 information; Bard, Bakeman, Boysen, and Leavens (2014) showed that human-reared
96 chimpanzees excel at social cognition. Van Leeuwen et al. (2021) found that chimpanzee
97 groups with higher social tolerance (measured via co-feeding proximity) were more likely to
98 act prosocially. Another line of research focused on personality traits (Altschul, Wallace,
99 Sonnweber, Tomonaga, & Weiss, 2017). For example, human-rated dominance and openness
100 to experiences correlated with problem-solving abilities (Hopper et al., 2014) and
101 Extraversion and agreeableness with sensitivity to inequity (Brosnan et al., 2015; Carter,
102 Marshall, Heinsohn, & Cowlishaw, 2014). Yet, personality is itself a latent, psychological
103 variable and the experiences that shape differences in personality remain unclear.

104 To summarize: studies on individual differences in great apes are promising, but rare.
105 One reason for this shortage is the difficulty of precise individual-level measurement
106 (Boogert, Madden, Morand-Ferron, & Thornton, 2018; Matzel & Sauce, 2017). To explore
107 cognitive structures or link abilities to external variables, reliable measures are essential. Yet,
108 reliability is rarely assessed in primate cognition research (Griffin, Guillette, & Healy, 2015).
109 For instance, the reliability of the widely used PCTB has yet to be systematically evaluated.

110 An exception is the work by Bohn et al. (2023). They combined several approaches to
111 studying individual differences while simultaneously assessing measurement quality. Over
112 two years, they tested individuals from all four great ape species on a variety of cognitive
113 tasks. They found that most – but not all – tasks reliably measured individual differences.
114 Stable cognitive differences were linked to long-term differences in experiences. However, due
115 to the small number of tasks, this study offered only limited insights into the structure of
116 individual differences.

117 The present study builds on Bohn et al. (2023) by addressing two key gaps. First, we

¹¹⁸ broadened the range of cognitive domains studied, including social cognition, numerical
¹¹⁹ reasoning, executive functions, and inferential reasoning. This approach allows us to test
¹²⁰ whether their findings replicate within these domains and generalize to others. Second, by
¹²¹ pooling data from both studies, we explored the correlations between cognitive traits within
¹²² and across domains, providing a deeper analysis of the structure of great ape cognition.

¹²³

Methods

¹²⁴ Participants

¹²⁵ A total of 48 great apes participated at least in one tasks at one time point. This
¹²⁶ included 12 Bonobos (*pan paniscus*, 4 females, age 3.60 to 40.70 years), 24 Chimpanzees (*pan*
¹²⁷ *troglodytes*, 17 females, age 3.80 to 57.80 years), 6 Gorillas (*gorilla gorilla*, 4 females, age 4.40
¹²⁸ to 24.40 years), and 6 Orangutans (*pongo abelii*, 5 females, age 4.70 to 43.10 years). The
¹²⁹ sample size at the different time points ranged from 34 to 45 for the different species (see
¹³⁰ supplementary material for details). All apes participated in cognitive research on a regular
¹³¹ basis. Apes were housed at the Wolfgang Köhler Primate Research Center located in Zoo
¹³² Leipzig, Germany. They lived in groups, with one group per species and two chimpanzee
¹³³ groups (group A and B). Research was noninvasive and strictly adhered to the legal
¹³⁴ requirements in Germany. Animal husbandry and research complied with the European
¹³⁵ Association of Zoos and Aquaria Minimum Standards for the Accommodation and Care of
¹³⁶ Animals in Zoos and Aquaria as well as the World Association of Zoos and Aquariums
¹³⁷ Ethical Guidelines for the Conduct of Research on Animals by Zoos and Aquariums.
¹³⁸ Participation was voluntary, all food was given in addition to the daily diet, and water was
¹³⁹ available ad libitum throughout the study. The study was approved by an internal ethics
¹⁴⁰ committee at the Max Planck Institute for Evolutionary Anthropology.

¹⁴¹ **Procedure**

¹⁴² Apes were tested in familiar sleeping or observation rooms by a single experimenter.
¹⁴³ The basic setup comprised a sliding table positioned in front of a mesh or a clear plexiglas
¹⁴⁴ panel. The experimenter sat on a small stool and used an occluder to cover the table (see
¹⁴⁵ Figure 1).

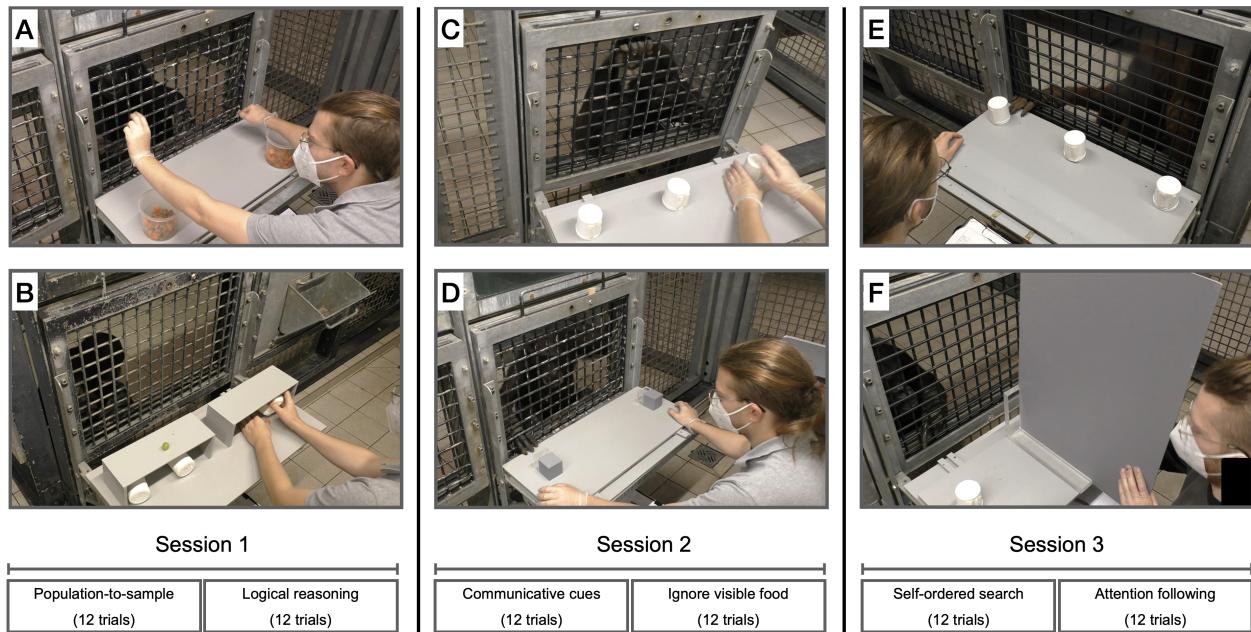


Figure 1. Setup used for the six tasks. A) population-to-sample, B) logical-reasoning, C) communicative-cues, D) ignore-visible-food, E) self-ordered-search and F) attention-following. Text at the bottom shows order of task presentation and trial numbers

¹⁴⁶ The study involved a total of six cognitive tasks. These were based on published
¹⁴⁷ procedures in the field of comparative psychology. The original publications often include
¹⁴⁸ control conditions to rule out alternative, cognitively less demanding ways to solve the tasks.
¹⁴⁹ We did not include such controls here and only ran the experimental conditions. For each
¹⁵⁰ task, we refer to these papers if they want to know more about control conditions and/or a
¹⁵¹ detailed discussion of the nature of the underlying cognitive mechanisms. Example videos for
¹⁵² each task can be found in the associated online repository. In the following, we give a brief
¹⁵³ description of each task. Additional details can be found in the supplementary material.

154 **Attention-following.** The Attention-following task was loosely modeled after

155 Kaminski, Call, and Tomasello (2004). The setup consisted of two identical cups placed on
156 the sliding table and a large opaque screen that was longer than the width of the sliding
157 table (Supplementary Figure 1F). The experimenter placed both cups on the table and
158 showed the ape that they were empty. Then, the experimenter baited both cups in view of
159 the ape and placed the opaque screen in the center between the two cups, perpendicular to
160 the mesh. Next, the experimenter moved to one side and looked at the cup in front of them.
161 Then, the experimenter pushed the sliding table forward and the ape was allowed to choose
162 one of the cups by pointing at it. If the ape chose the cup that the experimenter was looking
163 at, they received the food item. If they choose the other cup, they did not. We coded
164 whether the ape chose the side the experimenter was looking at (correct choice) or not. Apes
165 received twelve trials. The side at which the experimenter looked was counterbalanced with
166 same number of looks to each side and looks to the same side not more than two times in a
167 row. We assumed that apes follow the experimenters focus of attention to determine whether
168 or not their request could be seen and thus be successful.

169 **Communicative-cues.** This task was modeled after Schmid, Karg, Perner, and

170 Tomasello (2017). Three identical cups were placed equidistantly on a sliding table directly
171 in front of the ape (Figure 1C). In the beginning of a trial, the experimenter showed the ape
172 that all cups are empty. After placing an occluder between the subject and the cups, the
173 experimenter held up one food item and moved it behind the occluder, visiting all three cups
174 but baiting only one. Next, the occluder was lifted and E looked at the ape (ostensive cue),
175 called the name, and looked at one of the cups, while holding on to it with one hand and
176 tapping it with the other (continuous looking, 3 times tapping). Finally, the experimenter
177 pushed the sliding table forward for the ape to make a choice. If the ape chose the baited
178 cup, they received the reward – if not, not. We coded as correct choice if the ape chose the
179 indicated cup. Apes received twelve trials. The location of the indicated cup was
180 counterbalanced, with each cup being the target equally often and the same target not more

181 than two times in a row. We assumed that apes use the experimenter's communicative cues
182 to determine where the food is hidden.

183 **Ignore-visible-food.** The task was modeled after Völter, Tinklenberg, Call, and
184 Seed (2022). The task involved two opaque cups with an additional, sealed but transparent,
185 compartment attached to the front of each cup (facing the ape). For one cup, the
186 compartment contained a preferred food item that was clearly visible, for the other cup, the
187 compartment was empty (Figure 1D). In the beginning of the trial, the two cups were placed
188 upside down on the sliding table so that the ape could see that the opaque compartments of
189 both cups were empty. Next, the experimenter baited one of the cups in full view of the
190 subject. In non-conflict trials, the baited cup was the cup with the food item in the
191 transparent compartment. In conflict trials, the baited cup was the cup with the empty
192 compartment. After baiting the experimenter pushed the sliding table forwards and the ape
193 could chose by pointing. If the baited cup was chosen, the ape received the food. Apes
194 received 14 trials, twelve conflict trials and two non-conflict trials (1st and 8th trial). Only
195 conflict trials were analyzed. The location of the cup with the baited compartment was
196 counterbalanced, with the cup not being in the same location more than twice in a row. We
197 assumed that apes need to inhibit selecting the visible food item and instead use their
198 short-term memory to remember where the food was hidden.

199 **Logical-reasoning.** The task was modeled after Hanus and Call (2014). Three
200 identical cups were presented side-by-side on a sliding table, with the cup in the middle
201 sometimes positioned closer to the left cup and sometimes closer to the right.
202 (Supplementary Figure 1B). Two half-open boxes served as occluders to block the ape's view
203 when shuffling the cups. Each trial started by showing the ape that all three cups (one on
204 one side of the table, two on the other) were empty. After placing the occluders over both
205 sides of the table, the experimenter put one piece of food on top of each occluder. Next, the
206 experimenter hid each piece of food under the cup(s) behind the occluders. In case of the
207 occluder with the two cups, the food was randomly placed under one of the two cups while

208 both cups were visited and even shuffled. Finally, both occluders were lifted and the table
209 pushed forwards, allowing the ape to choose one of the three cups, from which they then
210 received the content. We coded whether the ape chose the certain cup (i.e. the cup from the
211 side of the table with only one cup). Apes received 12 trials. The side with one cup was
212 counterbalanced, with the same constellation appearing not more than two times in a row on
213 the same side. We assumed that apes would infer that the cup from the tray with only one
214 cup certainly contains food while the other cups contain food only in 50% of cases.

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

235 The crossing of the hands was also counterbalanced with no more than two crossings in a
236 row. We assumed that apes reasoned about the probability of the sample being a high
237 quality item based on observing the ratio in the population.

238 **Self-ordered-search.** The task was modeled after Völter, Mundry, Call, and Seed
239 (2019; Diamond, Prevor, Callender, and Druin, 1997; see also Petrides, 1995). Three
240 identical cups were placed equidistantly on a sliding table directly in front of the ape
241 (Supplementary Figure 1E). The experimenter baited all three cups in full view of the ape.
242 Next, the experimenter pushed the sliding table forwards for the ape to choose one of the
243 cups by pointing. After the choice, the table was pulled back and the ape received the food.
244 After a 3s pause, the table was pushed forward again for a second choice. This procedure
245 was repeated for a third choice. If the ape chose a baited cup, they received the food, if not,
246 not. We coded the number of times the ape chose an empty cup (i.e. chose a cup they
247 already chose before). Please note that this outcome variable differed from the other tasks in
248 two ways: first, possible values were 0, 1, and 2 (instead of just 0 and 1) and second, a lower
249 score indicated better performance. Apes received twelve trials. No counterbalancing was
250 needed. We assumed that apes use their working memory abilities to remember where they
251 had already searched and which cups still contained food.

252 **Predictor variables.** In addition to the data from the cognitive tasks, we collected
253 data for a range of predictor variables to predict individual differences in performance in the
254 cognitive tasks. Predictors could either vary with the individual (stable individual
255 characteristics: group, age sex, rearing history, and time spent in research), vary with
256 individual and time point (variable individual characteristics: rank, sickness, and sociality),
257 vary with group membership (group life: time spent outdoors, disturbances, and life events),
258 or vary with the testing arrangements and thus with individual, time point and session
259 (testing arrangements: presence of an observer, participation in other studies on the same
260 day or since the last time point). Predictors were collected from the zoo handbook with
261 demographic information about the apes, via a diary that the animal caretakers filled out on

262 a daily basis, or via proximity scans of the whole group. We provide a detailed description of
263 these variables in the supplementary material.

264 **Data collection**

265 Data collection started on April 28th, 2022, lasted until October 7th, 2023 and included
266 10 time points. One time point meant running all tasks with all participants. Within each
267 time point, the tasks were organized in three sessions (see Fig. 1). Session 1 included the
268 population-to-sample and logical-reasoning tasks, session 2 the communicative-cues and
269 ignore-visible-food tasks and session 3 the self-ordered-search and attention-following tasks.

270 The interval between two time points was planned to be eight weeks. However, it was
271 not always possible to follow this schedule so that some intervals were longer or shorter (see
272 supplementary material for details). The order of tasks was the same for all subjects. So was
273 the counterbalancing within each task. This exact procedure was repeated at each time point
274 so that the results would be comparable across participants and time points.

275 **Analysis, results and discussion**

276 To get an overview of the results, we first visualized the data (Fig. 2). Performance
277 was consistently above chance in the communicative-cues, ignore-visible-food and
278 population-to-sample tasks. For attention-following, this was the case only from time point 7
279 onward and for logical-reasoning, performance was, if anything, below chance. For the
280 self-ordered-search task, performance was below chance but here lower values reflect better
281 performance (i.e. systematic avoidance of the visible food item). For attention-following,
282 ignore-visible-food, communicative-cues and self-ordered-search there was a steady
283 improvement in performance over time.

284 In the following, we link performance in the tasks across time points to latent variables
285 representing cognitive abilities. We first ask how stable these abilities are over time and how
286 reliably they are measured. Next, we study the correlations between different abilities to

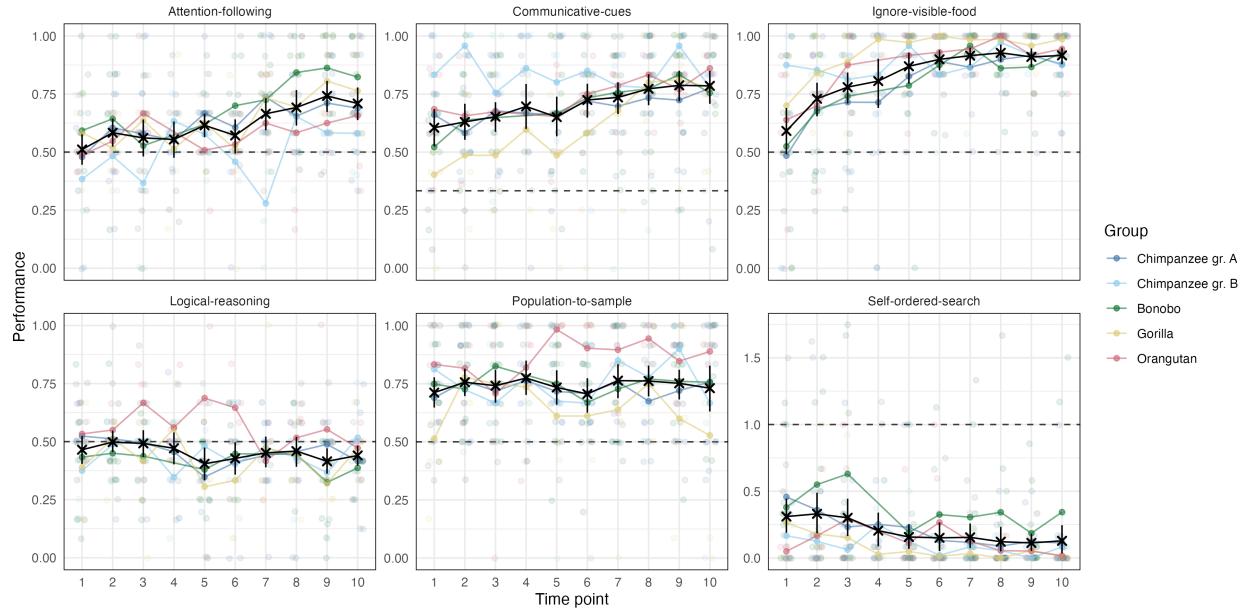


Figure 2. Results from the six cognitive tasks across time points. Black crosses show mean performance at each time point across species (with 95% CI). The sample size varied between time points and can be found in Supplementary Figure 1. Colored dots show mean performance by species. Dashed line shows chance level performance.

explore the internal structure of great ape cognition. Finally, we link performance in the tasks to external predictors to shed light on the sources of individual differences in abilities. Each section uses different statistical techniques which we describe in the respective section.

290 Stability and reliability

We first asked how robust performance was on a task-level, how stable individual differences were and how reliable the measures were. We used *Structural Equation Modeling* (SEM) (Bollen, 1989; Hoyle, 2012) to address these questions¹. For each task we fit three types of models that addressed different questions. We provide a detailed, mathematical

¹ SEMs usually use larger sample sizes than available in the present study. Bohn et al. (2023) reported a simulation study showing that parameters could be accurately estimated using Bayesian estimation techniques and reasonable model restrictions with sample sizes comparable to one we have here. We lay out the restrictive assumptions we imposed on the parameters in the supplementary material.

295 description of the models in the supplementary material.

296 We started with a latent state (LS) model. The goal of this model is to estimate a
297 measurement-error free latent state, representing an individual's cognitive ability, for each
298 time point. Measurement error is captured by dividing the trials from one time point into
299 two test-halves. Roughly speaking, the correlation between these two test-halves is an
300 indicator of measurement precision and used to estimate measurement error (and reliability).
301 Robustness of task-level performance can be assessed by comparing the means of latent
302 states across subjects for the different time points. Stability of individual differences can be
303 assessed by correlating latent states across different time points.

304 The temporal robustness of latent state means varied across tasks (Fig. 3A). In
305 attention-following, means increased over time and were significantly different from zero at
306 later time points (9 and 10). Communicative-cues and ignore-visible-food exhibited steady
307 increases, though ignore-visible-food saw a late-stage decline, with the latent mean at time
308 point 10 still significantly different from 0. Self-ordered-search showed a decrease (reduction
309 in errors) from time point 6 onward, while latent means for logical-reasoning and
310 population-to-sample remained stable throughout the study.

311 Correlations between latent states illustrated varying degrees of stability of individual
312 differences across tasks (Fig. 3B). Attention-following displayed low-to-moderate correlations
313 at early time points (before time point 7), increasing substantially thereafter.
314 Communicative-cues, ignore-visible-food, and self-ordered-search generally showed high
315 correlations between latent states (with time point 1 of ignore-visible-food being an
316 exception). Population-to-sample correlations were consistently high, while logical-reasoning
317 showed generally low, sometimes even negative, correlations, suggesting no stability across
318 time points.

319 Next, we fit two types of latent state-trait (LST) models. In comparison to the LS

models, these models assume that there is a single latent trait, representing an individual's stable cognitive ability, that is the same across time points. This way we can partition variation in performance on a given time point into variance due to the trait (consistency), variance due to the occasion (occasion specificity; 1 - consistency), and measurement error (used to estimate reliability). Like the latent states in the LS model, the trait in the LST model is assumed to be measurement error free (Geiser, 2020; Steyer, Ferring, & Schmitt, 1992; Steyer, Mayer, Geiser, & Cole, 2015). The first LST model we fit assumed that neither the absolute trait values nor the ranking of individuals changes over time (fixed means). This is the classic version of an LSTM. The second model allowed the absolute trait values to change over time while the ranking of individuals was fixed (varying means). Change over time according to this model is thus seen as change that is the same for all individuals. In both cases, stability of individual differences can be assessed by the proportion of variance explained by the trait (consistency).

Consistency estimates varied across tasks (Fig. 3C). In attention-following, the consistency coefficient was estimated to be 0.92 [Jana: possible to add 95%CrI?] for the fixed means model and 0.95 for the varying means model, indicating that more than 90% of true inter-individual differences were attributable to stable traits. However, given the low reliability of measurement (see below), this result should be interpreted with caution. [Jana: expand what that means]. For communicative-cues, consistency estimates differed the most between models and were higher in the varying means model (0.76) compared to the fixed means model (0.64). The reasons for this discrepancy is most likely the substantial change in mean performance over time in the task (see Fig. 3A). Ignore-visible-food showed similar consistency across models, with values of 0.59 (fixed means) and 0.64 (varying means). Logical-reasoning showed a similiar pattern to attention-following: Consistency was estimated to be high (fixed means: 0.81; varying means: 0.80) but reliability was low so that the same restrictions for interpretation apply. Self-ordered-search and population-to-sample had high consistency estimates according to both models: 0.79 (fixed means) and 0.84 (varying means)

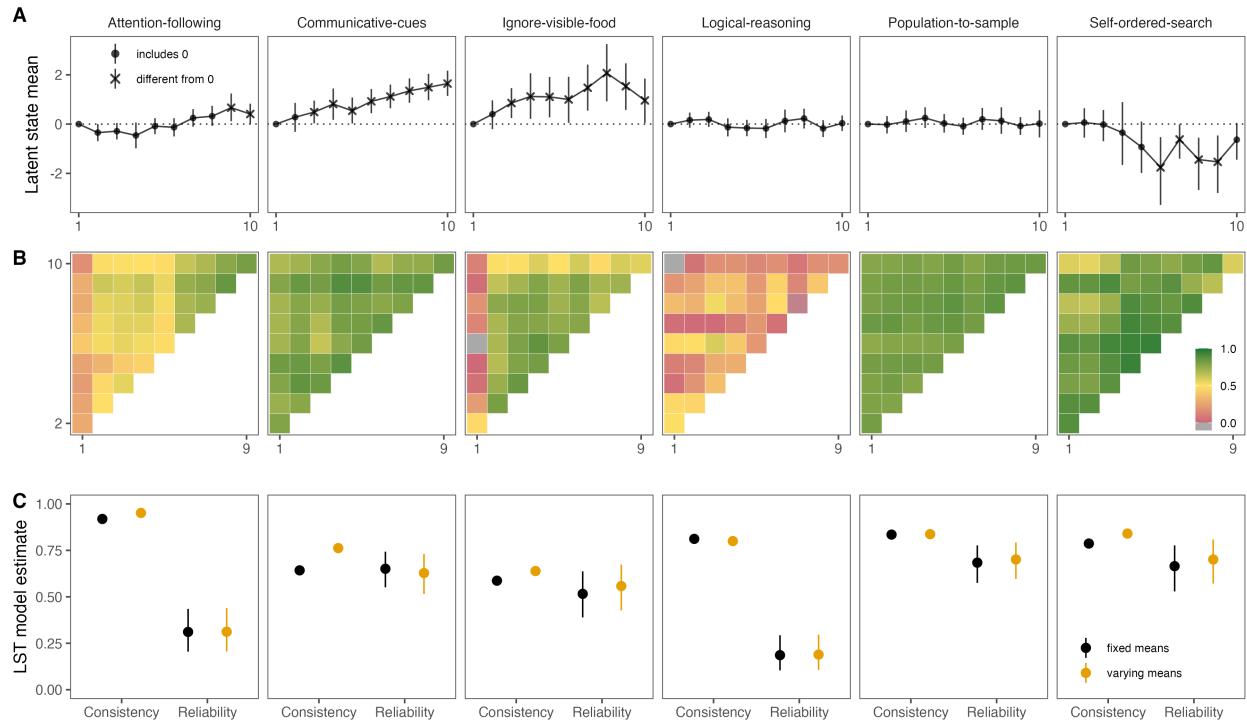


Figure 3. A) Latent mean estimates for each time point by task based on latent state model. Means at time point 1 are set to zero. Shape denotes whether the 95% CrI included zero (dashed line). The sample size varied between time points and can be found in Supplementary Fig. 1. B) Correlations between subject-level latent state estimates for the different time points by task. C) Mean estimates from latent state-trait models with fixed and varying means (color coded) with 95% CrI. Consistency refers to the proportion of (measurement-error-free) variance in performance explained by stable trait differences. Reliability refers to the proportion of true score variance to variance in raw scores.

347 for self-ordered-search and 0.83 (fixed means) and 0.84 (varying means) for
348 population-to-sample.

349 Reliability of measurement also varied significantly across tasks, based on the LST
350 models (Fig. 3C). For attention-following, reliability was initially low (fixed means: 0.31;
351 varying means: 0.31), but was substantially higher when only considering time points 7 and
352 onward (see supplementary material). Communicative-cues showed moderate reliability
353 (fixed means: 0.65; varying means: 0.63). Ignore-visible-food also had moderate reliability
354 across time points (fixed means: 0.52; varying means: 0.56). As mentioned above,
355 logical-reasoning exhibited very low reliability (fixed means: 0.19; varying means: 0.19).
356 Population-to-sample showed acceptable reliability (fixed means: 0.68; varying means: 0.70).
357 Self-ordered-search also exhibited acceptable reliability levels (fixed means: 0.66; varying
358 means: 0.70).

359 To summarize the SEM results, we saw that the six tasks differed substantially in what
360 they revealed about group- and individual-level variation. What stands out is the
361 widespread change in performance over time. For all tasks except population-to-sample and
362 logical-reasoning we observed an improvement in performance over time. This group-level
363 change, however, has different individual-level interpretations for the different tasks. For
364 communicative-cues, ignore-visible-food and self-ordered-search, individual differences
365 remained relatively stable despite the group-level change suggesting stable individual
366 differences combined with a systematic learning effect across individuals. In contrast, for
367 attention-following, there was little stability in individual differences at earlier time points
368 and only towards the end emerged a more stable ordering of individuals. In combination
369 with the low reliability at earlier time points, this suggests that at least some individuals
370 changed their response strategy in the course of the study. The combination of low reliability,
371 chance-level performance and low correlation of latent states for logical-reasoning suggests
372 that this task is not suited to assess individual differences in logical reasoning abilities in

373 great apes. It is also noteworthy that the reliability estimates are on average lower compared
374 to a previous study testing the same individuals on different tasks (Bohn et al., 2023). One
375 explanation might be the increase in performance over time. At the beginning of the study,
376 more individuals might have chosen randomly instead of using the available information
377 provided in the task setup and the demonstrations. By definition, random variation is not
378 reliable. With time, more and more individuals started using the available information so
379 that inter-individual differences in how good they are in using it could be detected.

380 Structure

381 To explore the structure of great ape cognition we correlated latent trait estimates for
382 each task. In contrast to raw performance scores, these estimates take into account the
383 reliability of measurement and are considered to be measurement-error free. Bohn et al.
384 (2023) tested the same individuals and we therefore also include the data from tasks reported
385 there (data from phase 2). Even though the data in the two studies was collected at different
386 time points, we think it is justifiable to analyse them jointly because the trait estimates
387 represent stable, time-invariant individual differences in cognitive abilities. The estimates
388 were computed ...

389 Figure 4 shows the correlations between trait estimates for the different tasks. Overall,
390 most correlations were not significantly different from zero (i.e. the 95% CI did include zero).
391 Because of this low average level of correlations, we decided not to explore models with
392 higher-order factors and will only interpret the qualitative patterns.

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

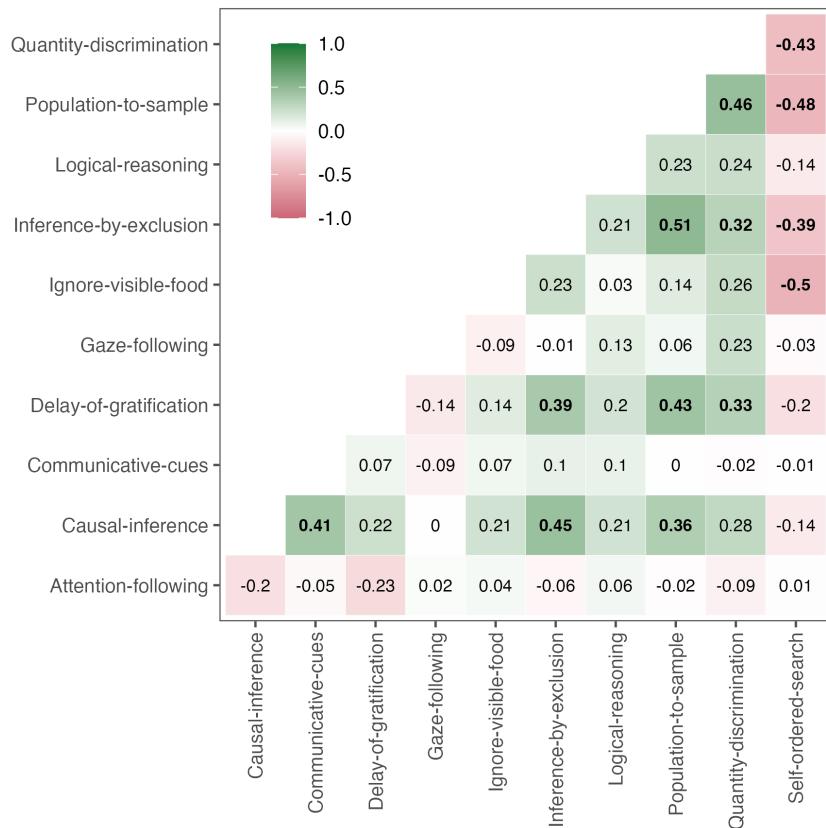


Figure 4. Correlations between trait estimates. Bold correlations have 95% CrI not overlapping with zero.

398 whether we find evidence for such a clustering in the data.

399 There was no significant correlation between any of the social cognition tasks.

400 Furthermore, attention-following and gaze-following did not correlate significantly with any
 401 of the other tasks and communicative-cues correlated only with causal-inference – a result we
 402 will discuss below. Thus, and in line with previous work (Herrmann et al., 2010), we found
 403 no evidence for shared cognitive processes in tasks measuring different aspects of social
 404 cognition.

405 The two tasks measuring reasoning about quantities did correlate significantly. Both
 406 tasks require discriminating between different quantities, directly in the case of

407 quantity-discrimination and as part of the decision making process in the case of
408 population-to-sample. Deciding between the samples from the two populations requires
409 discriminating between the relative quantities within each bucket from which the samples
410 were drawn.

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

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

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

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

447 Predictability

448 In this section, we analysed which external variables accounted for for inter- and
449 intra-individual differences in task performance. That is, we asked which of the predictor
450 variables described above predicted performance in the different tasks. Given the large
451 number of predictor variables (14), this question translates to a variable selection problem:
452 selecting a subset of variables from a larger pool. We used the projection predictive inference
453 (Piironen, Paasiniemi, & Vehtari, 2020) approach because it is a state-of-the-art procedure
454 that provides an excellent trade-off between model complexity and accuracy (Pavone,
455 Piironen, Bürkner, & Vehtari, 2020; Piironen & Vehtari, 2017). The projection prediction
456 approach is a two-step process: The first step consists of building the best predictive model
457 possible, called the reference model. In our case, the reference model is a Bayesian multilevel
458 regression model – fit via **brms** (Bürkner, 2017) – including all available predictors (Catalina,

459 Bürkner, & Vehtari, 2020). In the second step, the goal is to replace the posterior
460 distribution of the reference model with a simpler distribution containing fewer predictors
461 compared to the reference model. The importance of a predictor is assessed by inspecting
462 the mean log-predictive density (`elpd`) and root-mean-squared error (`rmse`) of models
463 containing the predictor vs. not.

464 The output of the procedure is a ranking of the different predictors. That is, for each
465 task, we get a ranking of how important a predictor is for constructing the simpler
466 replacement distribution. In addition, we can make a qualitatively assessment of whether or
467 not a predictor is relevant or not. In addition to the global assessment, we also inspected the
468 projected posterior distribution of the predictors classified as relevant to see how they
469 influenced performance. In the supplementary material we provide a detailed description of
470 the procedure including how the different variables were handled and how the importance of
471 each predictor was assessed.

472 In addition to the external predictors, the models also included a random intercept
473 term for subject ((`1 | subject`) in `brms` notation). This predictor was handled in a special
474 way in that it was always considered last because it would otherwise have soaked up most of
475 the variance before the other predictors would have had a chance to explain any of it.

476 Fig. 5A summarizes the selected predictors across tasks. For all tasks, the random
477 intercept term improved model fit the most (not shown in Fig. 5A). In line with results
478 reported by Bohn et al. (2023), this suggests that idiosyncratic developmental processes or
479 genetic pre-dispositions, which operate on time-scales longer than what we captured in our
480 study, accounted for a substantial portion of the variance in cognitive abilities between
481 individuals.

482 However, for two tasks, other predictors had an comparable explanatory power –
483 something that was not observe in Bohn et al. (2023). For population-to-sample, `time`

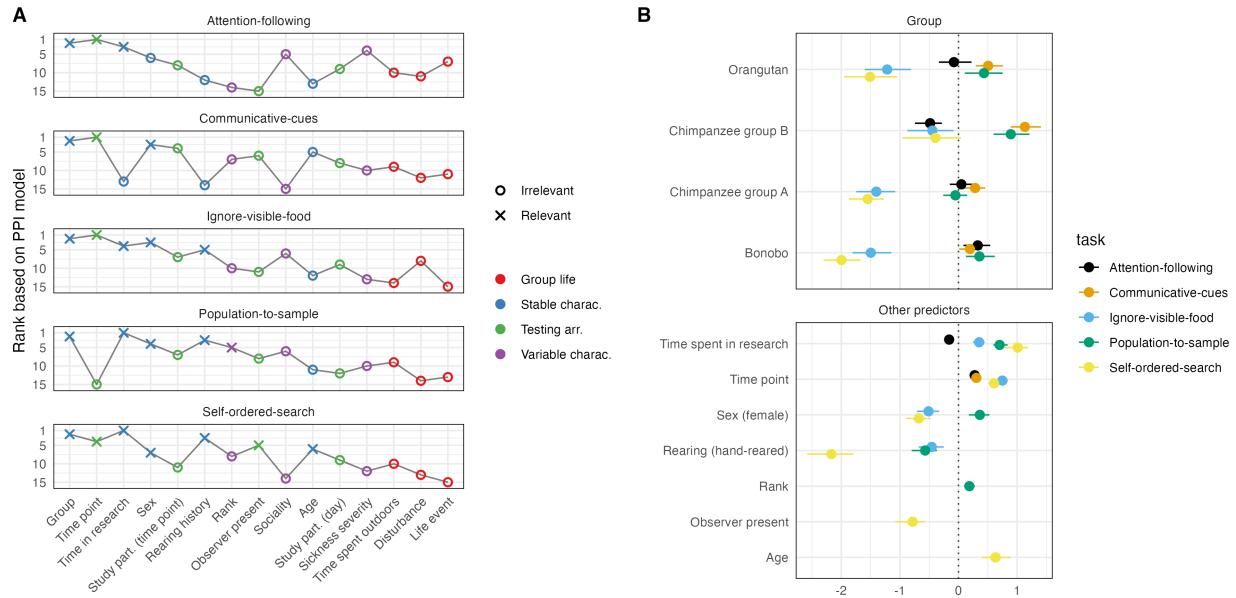


Figure 5. A) Predictor ranking and selection based on PPI models. Crosses mark predictors that were selected to be relevant based on the PPI models. Color shows the broader category each predictor belongs to. The x-axis is sorted by the average rank across tasks. B) Posterior model estimates for the selected predictors for each task based on data. Points show means with 95% Credible Interval. Color denotes task. For categorical predictors, the estimate gives the difference compared to the reference level (Gorilla for group).

484 spent in research improved the model fit even more than adding the random intercept at
 485 the end did. This could be interpreted that performance in this task strongly depends on
 486 having learned to pay attention to stimuli and the human experimenter. For
 487 ignore-visible-food, time point had an influence exceeding that of the random intercept
 488 term. We think this result reflects the strong within-task learning effect across subjects.
 489 Because performance increased substantially with time, most of the variation captured by
 490 time point exceeded the variation between individuals.

491 For the remaining predictors, the most highly-ranked and frequently selected ones
 492 came from the group of stable individual characteristics. The big exception being time
 493 point, which was ranked second across tasks. This pattern aligns with the SEM results, in

494 which we saw that most of the variance in performance could be traced back to stable trait
495 differences between individuals. The remaining occasion specific variation was largely due to
496 improvement over time, most likely reflecting task-specific learning processes. The remaining
497 time-varying predictors did not account for much variation.

498 The predictor selected most often was **group**. It was the only predictor that was
499 selected as relevant for all tasks. However, differences between groups were variable in that
500 the ranking of the groups changed from task to task (Fig. 5B). For example, Gorillas
501 performed best in ignore-visible-food and self-ordered-search, the Chimpanzee group B
502 performed best in communicative-cues and population-to-sample and the Bonobos performed
503 best in attention-following. This speaks against clear species or group differences in general
504 cognitive performance. Again, the most likely explanation for group differences is an
505 interaction between species specific dispositions and individual- / task-level developmental
506 processes.

507 The predictors that were selected more than once influenced performance in variable
508 ways (Fig. 5B). As mentioned above, **time point** always had a positive effect because
509 performance increased with time. Whenever **rearing** was selected to be relevant,
510 mother-reared individuals outperformed others. **Time spent in research** had a positive
511 effect, suggesting that more experience with research leads to better performance, except for
512 attention-following. The effect of **sex** was variable in that females outperformed males in
513 population-to-sample but males outperformed females in self-ordered-search and
514 ignore-visible-food.

515 General Discussion

516 In the present study, we investigated the stability, structure and predictability of great
517 ape cognition across a broad range of domains, including social cognition, reasoning about
518 quantities, executive functions, and inferential reasoning. We repeatedly administered six

519 tasks to a comparatively large sample of great apes a total of 10 times over a period of 1.5
520 years. Group-level results varied by task: while some tasks demonstrated substantial changes
521 over time, others remained relatively stable. The tasks also differed in how reliably they
522 measured individual differences, ranging from very poor (logical-reasoning) to very good
523 (population-to-sample, self-ordered-search). A significant portion of the observed variance in
524 performance could be attributed to stable differences in cognitive abilities between
525 individuals. However, these individual differences were not strongly associated across the full
526 range of tasks; instead smaller clusters of associations emerged. Finally, individual
527 differences in cognitive abilities were better predicted by stable, individual-specific
528 characteristics compared to transient aspects of everyday experience.

529 The observed substantial changes in performance over time highlight the plasticity of
530 cognition in great apes. Even though individual differences were stable – indicating that
531 individuals improved at similar rates – our findings show that adult apes, including older
532 individuals, are capable of learning and cognitive improvement. A case in point is the
533 chimpanzee B group, which consisted exclusively of adults, some of whom are in their 50s.
534 This contrasts with earlier work which suggested a decline in cognitive performance, in
535 particular executive functions, with age (Lacreuse, Parr, Chennareddi, & Herndon, 2018;
536 Lacreuse, Raz, Schmidtke, Hopkins, & Herndon, 2020; Manrique & Call, 2015). However,
537 earlier findings might have been driven by cohort effects in that longitudinal decline within
538 individuals was substantially smaller compared to cross-section differences between age
539 groups (Hopkins et al., 2021). In any case, this underscores the importance of longitudinal
540 studies to study the dynamics of cognitive development, not just early but also late in life.

541 The tasks substantially varied in their quality of measurement. This finding
542 emphasizes the importance of rigorously assessing measurement properties before including
543 tasks in cognitive test batteries or collecting data with large samples with the goals to assess
544 individual differences (see also Cauchoux et al., 2018; Soha, Peters, Anderson, Searcy, &

Nowicki, 2019). The reliability of measurement has profound implications for the conclusions that can be drawn. For instance, the logical-reasoning task showed no meaningful correlations with other tasks, which might suggest that logical reasoning is an isolated cognitive ability. However, the lack of correlation was more likely due to the task failing to measure anything reliably and variation in performance being predominantly noise.

We found no evidence for a *g*-factor explaining much of the differences between individuals (contra Hopkins et al., 2014). From a human perspective, and compared to earlier studies (Herrmann et al., 2010; Hopkins et al., 2014; Völter, Reindl, et al., 2022), the sample we tested could be considered small. However, we collected a large number of data points for each individual, and our analytical approach explicitly accounted for measurement reliability. This yielded robust estimates of individual-level cognitive abilities. Thus, we believe the lack of strong correlations across tasks reflects a genuine finding rather than noise. This pattern also aligns with previous work and animal cognition research more broadly (Poirier, Kozlovsky, Morand-Ferron, & Careau, 2020). For example, when conducting a confirmatory factor analysis on their data, Herrmann et al. (2010) less than half of the tasks in the PCTB loaded on any of the theoretically proposed factors and only 10 of the 105 bi-variate correlations were significant. Völter, Reindl, et al. (2022) found that only three out of 36 bi-variate correlations between executive functions tasks were significantly different from zero. Moving forward, perhaps a more fruitful approach would be to move away from a domain-level perspective to a process-level perspective. That is, instead of classifying tasks based on their domain of application (e.g. reasoning about the physical or social world), one should identify the cognitive processes involved in a task and generate predictions about correlations between tasks based on process-level commonalities. Case in point is the correlation observed between causal-inference and communicative-cues which can only be explained by a process-level perspective.

Finally, this study, alongside findings from Bohn et al. (2023), highlights that the

571 origins of individual differences in cognitive abilities most likely lie deeply embedded in the
572 ontogenetic history of individuals. Efforts to explain these differences by using easily
573 measurable variables, such as age, sex, or rank, were not fruitful. From this category, group
574 was the only predictor selected to be relevant for all tasks. It is important to note that group
575 is not the same as species in the present study: the two chimpanzee groups differed
576 substantially across tasks. This finding highlights the importance of studying within-species
577 variation and not only focusing on between-species variation. In and of itself, the group
578 variable has limited explanatory power because it encapsulates a variety of factors, including
579 age, social dynamics and genetic differences. Taken together, this highlights that to truly
580 understand the developmental origins of individual differences, longitudinal studies beginning
581 as early in life as possible are essential.

582

References

- 583 Altschul, D. M., Wallace, E. K., Sonnweber, R., Tomonaga, M., & Weiss, A. (2017).
584 Chimpanzee intellect: Personality, performance and motivation with touchscreen tasks.
585 *Royal Society Open Science*, 4(5), 170169.
- 586 Amici, F., Aureli, F., & Call, J. (2008). Fission-fusion dynamics, behavioral flexibility, and
587 inhibitory control in primates. *Current Biology*, 18(18), 1415–1419.
- 588 Bard, K. A., Bakeman, R., Boysen, S. T., & Leavens, D. A. (2014). Emotional engagements
589 predict and enhance social cognition in young chimpanzees. *Developmental Science*,
590 17(5), 682–696.
- 591 Beran, M. J. (2015). The comparative science of “self-control”: What are we talking about?
592 *Frontiers in Psychology*, 6, 51.
- 593 Beran, M. J., & Hopkins, W. D. (2018). Self-control in chimpanzees relates to general
594 intelligence. *Current Biology*, 28(4), 574–579.
- 595 Berdugo, S., Cohen, E., Davis, A., Matsuzawa, T., & Carvalho, S. (2023). Stable long-term
596 individual variation in chimpanzee technological efficiency. *bioRxiv*, 2023–2011.
- 597 Berio, L., & Moore, R. (2023). Great ape enculturation studies: A neglected resource in
598 cognitive development research. *Biology & Philosophy*, 38(2), 17.
- 599 Bohn, M., Eckert, J., Hanus, D., Lugauer, B., Holtmann, J., & Haun, D. B. (2023). Great
600 ape cognition is structured by stable cognitive abilities and predicted by developmental
601 conditions. *Nature Ecology & Evolution*, 7(6), 927–938.
- 602 Bollen, K. A. (1989). *Structural equations with latent variables* (Vol. 210). John Wiley &
603 Sons.
- 604 Boogert, N. J., Madden, J. R., Morand-Ferron, J., & Thornton, A. (2018). Measuring and
605 understanding individual differences in cognition. *Philosophical Transactions of the Royal
606 Society B: Biological Sciences*, 373(1756), 20170280.
- 607 Brosnan, S. F., Hopper, L. M., Richey, S., Freeman, H. D., Talbot, C. F., Gosling, S. D., ...
608 Schapiro, S. J. (2015). Personality influences responses to inequity and contrast in

- 609 chimpanzees. *Animal Behaviour*, 101, 75–87.
- 610 Burkart, J. M., Schubiger, M. N., & Schaik, C. P. van. (2017). The evolution of general
611 intelligence. *Behavioral and Brain Sciences*, 40.
- 612 Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan.
613 *Journal of Statistical Software*, 80(1), 1–28.
- 614 Carter, A. J., Marshall, H. H., Heinsohn, R., & Cowlishaw, G. (2014). Personality predicts
615 the propensity for social learning in a wild primate. *PeerJ*, 2, e283.
- 616 Catalina, A., Bürkner, P.-C., & Vehtari, A. (2020). Projection predictive inference for
617 generalized linear and additive multilevel models. *arXiv Preprint arXiv:2010.06994*.
- 618 Cauchoix, M., Chow, P., Van Horik, J., Atance, C., Barbeau, E., Barragan-Jason, G., et
619 al.others. (2018). The repeatability of cognitive performance: A meta-analysis.
620 *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1756),
621 20170281.
- 622 Deaner, R. O., Van Schaik, C. P., & Johnson, V. (2006). Do some taxa have better
623 domain-general cognition than others? A meta-analysis of nonhuman primate studies.
624 *Evolutionary Psychology*, 4(1), 147470490600400114.
- 625 Diamond, A., Prevor, M. B., Callender, G., & Druin, D. P. (1997). Prefrontal cortex
626 cognitive deficits in children treated early and continuously for PKU. *Monographs of the*
627 *Society for Research in Child Development*, i–206.
- 628 Dunbar, R., & Shultz, S. (2017). Why are there so many explanations for primate brain
629 evolution? *Philosophical Transactions of the Royal Society B: Biological Sciences*,
630 372(1727), 20160244.
- 631 Eckert, J., Call, J., Hermes, J., Herrmann, E., & Rakoczy, H. (2018). Intuitive statistical
632 inferences in chimpanzees and humans follow weber's law. *Cognition*, 180, 99–107.
- 633 Fichtel, C., Dinter, K., & Kappeler, P. M. (2020). The lemur baseline: How lemurs compare
634 to monkeys and apes in the primate cognition test battery. *PeerJ*, 8, e10025.
- 635 Fröhlich, M., Wittig, R. M., & Pika, S. (2019). The ontogeny of intentional communication

- 636 in chimpanzees in the wild. *Developmental Science*, 22(1), e12716.
- 637 Geiser, C. (2020). *Longitudinal structural equation modeling with mplus: A latent state-trait*
638 *perspective*. Guilford Publications.
- 639 Griffin, A. S., Guillette, L. M., & Healy, S. D. (2015). Cognition and personality: An
640 analysis of an emerging field. *Trends in Ecology & Evolution*, 30(4), 207–214.
- 641 Hanus, D., & Call, J. (2014). When maths trumps logic: Probabilistic judgements in
642 chimpanzees. *Biology Letters*, 10(12), 20140892.
- 643 Henke-von der Malsburg, J., Kappeler, P. M., & Fichtel, C. (2020). Linking ecology and
644 cognition: Does ecological specialisation predict cognitive test performance? *Behavioral*
645 *Ecology and Sociobiology*, 74(12), 154.
- 646 Herrmann, E., Call, J., Hernández-Lloreda, M. V., Hare, B., & Tomasello, M. (2007).
647 Humans have evolved specialized skills of social cognition: The cultural intelligence
648 hypothesis. *Science*, 317(5843), 1360–1366.
- 649 Herrmann, E., Hernández-Lloreda, M. V., Call, J., Hare, B., & Tomasello, M. (2010). The
650 structure of individual differences in the cognitive abilities of children and chimpanzees.
651 *Psychological Science*, 21(1), 102–110.
- 652 Hopkins, W. D., Mareno, M. C., Neal Webb, S. J., Schapiro, S. J., Raghanti, M. A., &
653 Sherwood, C. C. (2021). Age-related changes in chimpanzee (*pan troglodytes*) cognition:
654 Cross-sectional and longitudinal analyses. *American Journal of Primatology*, 83(3),
655 e23214.
- 656 Hopkins, W. D., Russell, J. L., & Schaeffer, J. (2014). Chimpanzee intelligence is heritable.
657 *Current Biology*, 24(14), 1649–1652.
- 658 Hopper, L. M., Price, S. A., Freeman, H. D., Lambeth, S. P., Schapiro, S. J., & Kendal, R. L.
659 (2014). Influence of personality, age, sex, and estrous state on chimpanzee
660 problem-solving success. *Animal Cognition*, 17, 835–847.
- 661 Hoyle, R. H. (2012). *Handbook of structural equation modeling*. Guilford press.
- 662 Joly, M., Micheletta, J., De Marco, A., Langermans, J. A., Sterck, E. H., & Waller, B. M.

- 663 (2017). Comparing physical and social cognitive skills in macaque species with different
664 degrees of social tolerance. *Proceedings of the Royal Society B: Biological Sciences*,
665 284(1862), 20162738.
- 666 Kaigaishi, Y., Nakamichi, M., & Yamada, K. (2019). High but not low tolerance populations
667 of Japanese macaques solve a novel cooperative task. *Primates*, 60, 421–430.
- 668 Kaminski, J., Call, J., & Tomasello, M. (2004). Body orientation and face orientation: Two
669 factors controlling apes' begging behavior from humans. *Animal Cognition*, 7, 216–223.
- 670 Kaufman, A. B., Reynolds, M. R., & Kaufman, A. S. (2019). The structure of ape
671 (hominoidea) intelligence. *Journal of Comparative Psychology*, 133(1), 92.
- 672 Lacreuse, A., Parr, L., Chennareddi, L., & Herndon, J. G. (2018). Age-related decline in
673 cognitive flexibility in female chimpanzees. *Neurobiology of Aging*, 72, 83–88.
- 674 Lacreuse, A., Raz, N., Schmidtke, D., Hopkins, W. D., & Herndon, J. G. (2020). Age-related
675 decline in executive function as a hallmark of cognitive ageing in primates: An overview
676 of cognitive and neurobiological studies. *Philosophical Transactions of the Royal Society*
677 B, 375(1811), 20190618.
- 678 Manrique, H. M., & Call, J. (2015). Age-dependent cognitive inflexibility in great apes.
679 *Animal Behaviour*, 102, 1–6.
- 680 ManyPrimates, Aguenounon, G., Allritz, M., Altschul, D. M., Ballesta, S., Beaud, A., et
681 al.others. (2022). The evolution of primate short-term memory. *Animal Behavior and*
682 *Cognition*, 9(4), 428–516.
- 683 Matzel, L. D., & Sauce, B. (2017). Individual differences: Case studies of rodent and primate
684 intelligence. *Journal of Experimental Psychology: Animal Learning and Cognition*, 43(4),
685 325.
- 686 Pavone, F., Piironen, J., Bürkner, P.-C., & Vehtari, A. (2020). *Using reference models in*
687 *variable selection*. Retrieved from <https://arxiv.org/abs/2004.13118>
- 688 Petrides, M. (1995). Impairments on nonspatial self-ordered and externally ordered working
689 memory tasks after lesions of the mid-dorsal part of the lateral frontal cortex in the

- monkey. *Journal of Neuroscience*, 15(1), 359–375.
- Piantadosi, S. T., & Kidd, C. (2016). Extraordinary intelligence and the care of infants. *Proceedings of the National Academy of Sciences*, 113(25), 6874–6879.
- 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. <https://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.
<https://doi.org/10.1007/s11222-016-9649-y>
- Poirier, M.-A., Kozlovsky, D. Y., Morand-Ferron, J., & Careau, V. (2020). How general is cognitive ability in non-human animals? A meta-analytical and multi-level reanalysis approach. *Proceedings of the Royal Society B*, 287(1940), 20201853.
- Reader, S. M., Hager, Y., & Laland, K. N. (2011). The evolution of primate general and cultural intelligence. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1567), 1017–1027.
- Rosati, A. G. (2017). Foraging cognition: Reviving the ecological intelligence hypothesis. *Trends in Cognitive Sciences*, 21(9), 691–702.
- Rosati, A. G., & Santos, L. R. (2017). Tolerant barbary macaques maintain juvenile levels of social attention in old age, but despotic rhesus macaques do not. *Animal Behaviour*, 130, 199–207.
- Schmid, B., Karg, K., Perner, J., & Tomasello, M. (2017). Great apes are sensitive to prior reliability of an informant in a gaze following task. *PLoS One*, 12(11), e0187451.
- Schmitt, V., Pankau, B., & Fischer, J. (2012). Old world monkeys compare to apes in the primate cognition test battery. *PloS One*, 7(4), e32024.
- Schubiger, M. N., Fichtel, C., & Burkart, J. M. (2020). Validity of cognitive tests for non-human animals: Pitfalls and prospects. *Frontiers in Psychology*, 11, 1835.
- Shaw, R. C., & Schmelz, M. (2017). Cognitive test batteries in animal cognition research:

- 717 Evaluating the past, present and future of comparative psychometrics. *Animal Cognition*,
718 20(6), 1003–1018.
- 719 Sih, A., Sinn, D. L., & Patricelli, G. L. (2019). On the importance of individual differences
720 in behavioural skill. *Animal Behaviour*, 155, 307–317.
- 721 Soha, J. A., Peters, S., Anderson, R. C., Searcy, W. A., & Nowicki, S. (2019). Performance
722 on tests of cognitive ability is not repeatable across years in a songbird. *Animal*
723 *Behaviour*, 158, 281–288.
- 724 Steyer, R., Ferring, D., & Schmitt, M. J. (1992). States and traits in psychological
725 assessment. *European Journal of Psychological Assessment*, 8, 79–98.
- 726 Steyer, R., Mayer, A., Geiser, C., & Cole, D. A. (2015). A theory of states and
727 traits—revised. *Annual Review of Clinical Psychology*, 11, 71–98.
- 728 Thornton, A., & Lukas, D. (2012). Individual variation in cognitive performance:
729 Developmental and evolutionary perspectives. *Philosophical Transactions of the Royal*
730 *Society B: Biological Sciences*, 367(1603), 2773–2783.
- 731 Van Leeuwen, E. J., DeTroy, S. E., Kaufhold, S. P., Dubois, C., Schütte, S., Call, J., & Haun,
732 D. B. (2021). Chimpanzees behave prosocially in a group-specific manner. *Science*
733 *Advances*, 7(9), eabc7982.
- 734 Völter, C. J., Mundry, R., Call, J., & Seed, A. M. (2019). Chimpanzees flexibly update
735 working memory contents and show susceptibility to distraction in the self-ordered search
736 task. *Proceedings of the Royal Society B*, 286(1907), 20190715.
- 737 Völter, C. J., Reindl, E., Felsche, E., Civelek, Z., Whalen, A., Lugosi, Z., ... Seed, A. M.
738 (2022). The structure of executive functions in preschool children and chimpanzees.
739 *Scientific Reports*, 12(1), 1–16.
- 740 Völter, C. J., Tinklenberg, B., Call, J., & Seed, A. M. (2018). Comparative psychometrics:
741 Establishing what differs is central to understanding what evolves. *Philosophical*
742 *Transactions of the Royal Society B: Biological Sciences*, 373(1756), 20170283.
- 743 Völter, C. J., Tinklenberg, B., Call, J., & Seed, A. M. (2022). Inhibitory control and cue

744 relevance modulate chimpanzees' (*pan troglodytes*) performance in a spatial foraging task.

745 *Journal of Comparative Psychology*, 136(2), 105.

746 Watson, S. K., Vale, G. L., Hopper, L. M., Dean, L. G., Kendal, R. L., Price, E. E., et

747 al.others. (2018). Chimpanzees demonstrate individual differences in social information

748 use. *Animal Cognition*, 21, 639–650.