

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

³ Manuel Bohn^{1,2}, Christoph J. Völter^{2,3}, Daniel Hanus², Nico Eisbrenner², Johanna Eckert²,
⁴ Jana Holtmann⁴, & 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 Comparative Cognition, Messerli Research Institute, University of Veterinary Medicine
⁹ Vienna, Medical University of Vienna and University of Vienna, Vienna, Austria

¹⁰ ⁴ Wilhelm Wundt Institute of Psychology, Leipzig University, Leipzig, Germany

11

Author Note

12 Manuel Bohn was funded by a Jacobs Foundation Research Fellowship (Grant
13 No. 2022-1484-00) and a Lower Saxony Impulse Professorship through the
14 zukunft.niedersachsen program. We thank the Max Planck Society for the Advancement of
15 Science.

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

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

27

Abstract

28 Variation in cognitive abilities is critical to understanding both the evolution and
29 development of cognition. In this study, we examined the stability, structure, and
30 predictability of individual differences in cognitive abilities in great apes across a broad
31 range of domains, including social cognition, reasoning about quantities, executive functions,
32 and inferential reasoning. We repeatedly administered six established tasks to N = 48 apes
33 from four great ape species, spanning 10 sessions over 1.5 years. Task performance was most
34 strongly predicted by stable, individual-specific characteristics rather than transient or
35 group-level variables, highlighting the need for ontogenetic studies to understand cognitive
36 variation in great apes. Furthermore, there were substantial correlations between tasks:
37 associations between all non-social tasks were large and positive, suggesting shared cognitive
38 processes. In contrast, tasks measuring social cognition were neither correlated with each
39 other nor with non-social measures. Future studies of great ape cognition should build
40 mechanistic models of cognitive processes to build an understanding of the evolution of
41 cognition based on process-level commonalities across species.

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

43 Individual differences in great ape cognition across time and domains: stability, structure,
44 and predictability

45 **Introduction**

46 Variation fuels evolution. Individual differences in cognitive abilities are essential for
47 understanding what evolves [1–3]. These differences reveal which aspects of cognition are
48 invariant and which are malleable. They also shed light on the broader structure of the
49 cognitive architecture by identifying relationships between different cognitive abilities.
50 Moreover, they help identify the socio-ecological factors shaping cognition during both
51 ontogeny and phylogeny.

52 Broadly speaking, great ape cognition is marked by substantial individual variability
53 across functional domains, such as tool use, communication, social cognition, causal
54 reasoning, and reasoning about quantities. This variability has been observed in both captive
55 and wild settings [4–8] and suggests significant plasticity in cognitive abilities, presumably
56 shaped by social and ecological influences. As noted above, such individual differences can
57 be used to study the structure of great ape cognition and its origins [1].

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 [9–11]. Using the Primate Cognition Test Battery (PCTB) [12], [4]
61 found no evidence for a single *g*-factor in chimpanzees. Instead, they observed a bifactorial
62 structure, with one factor linked to spatial tasks and the other to social and physical tasks.
63 Similar findings have been reported for other primates [13,14]. By contrast, [15] used the
64 PCTB to test a different sample of chimpanzees and identified a *g*-factor, which was later
65 found to relate to measures of self-control [16]. However, this study did not test whether the
66 proposed structure (a single *g*-factor) fit the data well. In a subsequent re-analysis, [17]
67 combined datasets collected with the PCTB and found the single *g*-factor model inadequate.
68 Only multidimensional models accurately described the data. Beyond general cognitive

69 abilities, [18] investigated the structure of executive functions in chimpanzees using a
70 multi-trait, multi-method approach. Their results showed limited evidence for the structure
71 proposed for executive functions in humans.

72 The existence of individual differences raises questions about their origins. Most
73 theories about the factors influencing the emergence of complex cognitive abilities operate on
74 a species level [19–21]. Empirical studies in this tradition often compare closely related
75 species with differing social structures or ecological pressures [22–25]. Alternatively,
76 researchers aggregate data across studies to compare species on a larger scale [9,26]. This
77 approach, however, faces challenges in comparability, as data are often collected using
78 inconsistent methods [27]. An exception is [28,see also 29], which employed standardized
79 methods to collect a large dataset on short-term memory and test species-level hypotheses.
80 However, their results were surprising: no single socio-ecological predictor explained
81 cognitive variation beyond phylogenetic relatedness.

82 In contrast, much less research has focused on the individual level [30]. Early work
83 focused on the effects of enculturation—raising great apes in a human environment. Most of
84 these studies, however, involved only one individual, making it difficult to identify the
85 relevant aspects of experience that led to the observed changes in cognition [see 31 for a
86 recent summary]. Few studies with larger samples exist: [5] found that hand-reared
87 chimpanzees are more likely to use social information; [32] showed that human-reared
88 chimpanzees excel at social cognition. [33] found that chimpanzee groups with higher social
89 tolerance (measured via co-feeding proximity) were more likely to act prosocially. Another
90 line of research focused on personality traits [34]. For example, human-rated dominance and
91 openness to experience correlated with problem-solving abilities [35], and extraversion and
92 agreeableness with sensitivity to inequity [36,37]. Yet, personality is itself a latent
93 psychological variable, and the experiences that shape differences in personality remain
94 unclear.

95 To summarize: studies on individual differences in great apes are promising but rare.

96 One reason for this shortage is the difficulty of precise individual-level measurement [38,39].

97 To explore cognitive structures or link abilities to external variables, reliable measures are

98 essential. Nevertheless, reliability is rarely assessed in primate cognition research [40]. For

99 instance, the reliability of the widely used PCTB has yet to be systematically evaluated.

100 An exception is the work by [6]. They combined several approaches to studying

101 individual differences while simultaneously assessing measurement quality. Over two years,

102 they tested individuals from four great ape species on a variety of cognitive tasks. They

103 found that most—but not all—tasks reliably measured individual differences. Stable

104 cognitive differences were linked to long-term differences in experiences. However, due to the

105 small number of tasks, this study offered only limited insights into the structure of individual

106 differences.

107 The present study builds on [6] by addressing two key gaps. First, we broadened the

108 range of cognitive domains studied, including social cognition, reasoning about quantities,

109 executive functions, and inferential reasoning. This approach allows us to test whether their

110 findings replicate within these domains and generalize to others. Second, by pooling data

111 from both studies, we explored the correlations between cognitive traits within and across

112 domains, providing a deeper analysis of the structure of great ape cognition.

113 Methods

114 Participants

115 A total of 48 great apes participated at least in one tasks at one time point. This

116 included 12 Bonobos (*pan paniscus*, 4 females, age 3.60 to 40.70 years), 24 Chimpanzees (*pan*

117 *troglodytes*, 17 females, age 3.80 to 57.80 years), 6 Gorillas (*gorilla gorilla*, 4 females, age 4.40

118 to 24.40 years), and 6 Orangutans (*pongo abelii*, 5 females, age 4.70 to 43.10 years). The

119 sample size at the different time points ranged from 34 to 45 for the different species (see

supplementary material for details). All apes participated in cognitive research on a regular basis. 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.

Procedure

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

The study involved a total of six cognitive tasks. These were based on published procedures 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 these papers to learn more about control conditions and/or a detailed discussion of the nature of the presumed underlying cognitive mechanisms. Example videos for each task can be found in the associated online repository (<https://github.com/ccp-eva/laac2/tree/main/videos>). A second coder, unfamiliar to the purpose of the study, coded 20% of all time points for all tasks. Inter-rater reliability was excellent (lowest proportion of agreement = 0.99 for population-to-sample, lowest $\kappa = 0.97$

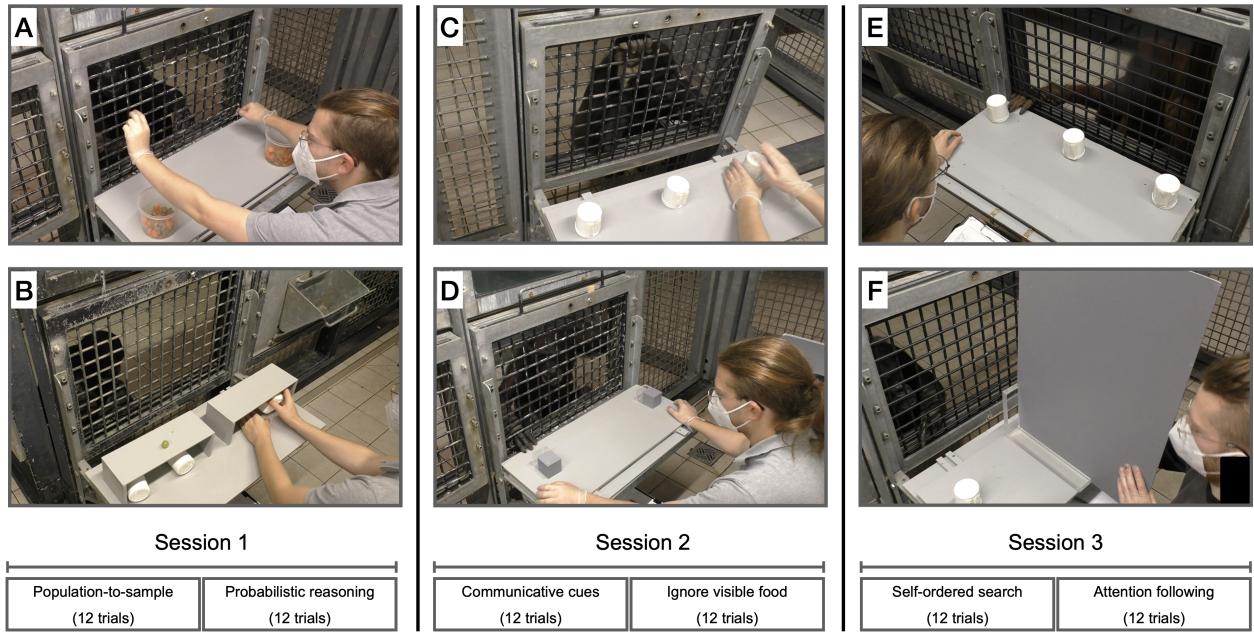


Figure 1. Setup used for the six tasks. A) population-to-sample, B) probabilistic-reasoning, C) communicative-cues, D) ignore-visible-food, E) self-ordered-search and F) attention-following. Text at the bottom shows order of task presentation and trial numbers

146 for ignore-visible-food). Additional details can be found in the supplementary material.

147 **Attention-following.** The Attention-following task was loosely modeled after [41].
 148 The setup consisted of two identical cups placed on the sliding table and a large opaque
 149 screen that was longer than the width of the sliding table (Supplementary Figure 1F). The
 150 experimenter placed both cups on the table and showed the ape that they were empty. Then,
 151 the experimenter baited both cups in view of the ape and placed the opaque screen in the
 152 center between the two cups, perpendicular to the mesh. Next, the experimenter moved to
 153 one side and looked at the cup in front of them. Then, the experimenter pushed the sliding
 154 table forward and the ape was allowed to choose one of the cups by pointing at it. If the ape
 155 chose the cup that the experimenter was looking at, they received the food item. If they
 156 chose the other cup, they did not. We coded whether the ape chose the side the
 157 experimenter was looking at (correct choice) or not. Apes received twelve trials. The side at
 158 which the experimenter looked was counterbalanced with the same number of looks to each

159 side and looks to the same side not more than two times in a row. We assumed that apes
160 follow the experimenter's focus of attention to determine whether or not their request could
161 be seen and thus be successful.

162 **Communicative-cues.** This task was modeled after [42]. Three identical cups were
163 placed equidistantly on a sliding table directly in front of the ape (Figure 1C). In the
164 beginning of a trial, the experimenter showed the ape that all cups are empty. After placing
165 an occluder between the subject and the cups, the experimenter held up one food item and
166 moved it behind the occluder, visiting all three cups but baiting only one. Next, the occluder
167 was lifted and E looked at the ape (ostensive cue), called the ape's name, and looked at one
168 of the cups, while holding on to it with one hand and tapping it with the other (continuous
169 looking, 3 times tapping). Finally, the experimenter pushed the sliding table forward for the
170 ape to make a choice. If the ape chose the baited cup, they received the reward – if not, not.
171 We coded as correct choice if the ape chose the indicated cup. Apes received twelve trials.
172 The location of the indicated cup was counterbalanced, with each cup being the target
173 equally often and the same target not more than two times in a row. We assumed that apes
174 use the experimenter's communicative cues to determine where the food is hidden.

175 **Ignore-visible-food.** The task was modeled after [43]. The task involved two
176 opaque cups with an additional, sealed but transparent, compartment attached to the front
177 of each cup (facing the ape). For one cup, the compartment contained a preferred food item
178 that was clearly visible, for the other cup, the compartment was empty (Figure 1D). In the
179 beginning of the trial, the two cups were placed upside down on the sliding table so that the
180 ape could see that the opaque compartments of both cups were empty. Next, the
181 experimenter baited one of the cups in full view of the subject. In non-conflict trials, the
182 baited cup was the cup with the food item in the transparent compartment. In conflict trials,
183 the baited cup was the cup with the empty compartment. After baiting the experimenter
184 pushed the sliding table forwards and the ape could choose by pointing. If the baited cup
185 was chosen, the ape received the food. Apes received 14 trials, twelve conflict trials and two

186 non-conflict trials (1st and 8th trial). Only conflict trials were analyzed. The location of the
187 cup with the baited compartment was counterbalanced, with the cup not being in the same
188 location more than two times in a row. We assumed that apes inhibit selecting the visible
189 food item and instead use their short-term memory to remember where the food was hidden.

190 **Probabilistic-reasoning.** The task was modeled after [44]. Three identical cups
191 were presented side-by-side on a sliding table, with the cup in the middle sometimes
192 positioned close to the left cup and sometimes close to the right. (Supplementary Figure 1B).
193 Two half-open boxes served as occluders to block the ape's view when shuffling the cups.
194 Each trial started by showing the ape that all three cups (one on one side of the table, two
195 on the other) were empty. After placing the occluders over both sides of the table, thereby
196 covering two cups on one side and one cup on the other, the experimenter put one piece of
197 food on top of each occluder. Next, the experimenter hid each piece of food under the cup(s)
198 behind the occluders. In case of the occluder with the two cups, the food was randomly
199 placed under one of the two cups while both cups were visited and even shuffled. Finally,
200 both occluders were lifted and the table pushed forwards, allowing the ape to choose one of
201 the three cups, from which they then received the content. We coded whether the ape chose
202 the certain cup (i.e. the cup from the side of the table with only one cup). Apes received 12
203 trials. The side with one cup was counterbalanced, with the same constellation appearing
204 not more than two times in a row on the same side. We assumed that apes would infer that
205 the cup from the tray with only one cup certainly contains food while the other cups contain
206 food only in 50% of cases.

207 **Population-to-sample.** The task was modeled after [45, see also 46]. During the test,
208 apes saw two transparent buckets filled with pellets and carrot pieces (the carrot pieces had
209 roughly the same size and shape as the pellets). Each bucket contained 80 food items. The
210 distribution of pellets to carrot pieces was 4:1 in bucket A, and 1:4 in bucket B. Pellets are
211 preferred food items in comparison to carrots. The experimenter placed both buckets on a
212 table, one left, one right (Figure 1A). In the beginning of a trial, the experimenter picked up

213 the bucket on the right side, tilted it forward so the ape could see inside, placed it back on
214 the table and turned it around 360°. The same procedure was repeated with the other bucket.
215 Next, the experimenter looked at the ceiling, inserted each hand in the bucket in front of it
216 and drew one item from the bucket without the ape seeing which type (E picked always of
217 the majority type). The food items remained hidden in the experimenter's fists. Next, the
218 experimenter extended the arms (in parallel) towards the ape who was then allowed to make
219 a choice by pointing to one of the fists. The ape received the chosen sample. In half of the
220 trials, the experimenter crossed arms when moving the fists towards the ape to ensure that
221 the apes made a choice between samples and not just chose the side where the favorable
222 population (bucket) was still visible. In between trials, the buckets were refilled to restore
223 the original distributions. Apes received twelve trials. We coded whether the ape chose the
224 sample from the population with the higher number of preferred food items. The location of
225 the buckets (left and right) was counterbalanced, with the buckets in the same location no
226 more than two times in a row. The crossing of the hands was also counterbalanced with no
227 more than two crossings in a row. We assumed that apes reasoned about the probability of
228 the sample being a preferred item based on observing the ratio in the population.

229 **Self-ordered-search.** The task was modeled after [47,see also 48,49]. Three
230 identical cups were placed equidistantly on a sliding table directly in front of the ape
231 (Supplementary Figure 1E). The experimenter baited all three cups in full view of the ape.
232 Next, the experimenter pushed the sliding table forwards for the ape to choose one of the
233 cups by pointing. After the choice, the table was pulled back and the ape received the food.
234 After a 3s pause, the table was pushed forward again for a second choice. This procedure
235 was repeated for a third choice. If the ape chose a baited cup, they received the food, if not,
236 not. We coded the number of times the ape chose an empty cup (i.e. chose a cup they
237 already chose before). Please note that this outcome variable differed from the other tasks in
238 two ways: first, possible values were 0, 1, and 2 (instead of just 0 and 1) and second, a lower
239 score indicated better performance. Apes received twelve trials. No counterbalancing was

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

Predictor variables. In addition to the data from the cognitive tasks, we collected data for a range of predictor variables to predict individual differences in performance in the cognitive tasks. Predictors could either vary with the individual (stable individual characteristics: group, age, sex, rearing history, and time spent in research), vary with individual and time point (variable individual characteristics: rank, sickness, and sociality), vary with group membership (group life: time spent outdoors, disturbances, and life events), or vary with the testing arrangements and thus with individual, time point and session (testing arrangements: presence of an observer, participation in other studies on the same day and since the last time point). Predictors were collected from the zoo handbook with demographic information about the apes, via a diary that the animal caretakers filled out on a daily basis, or via proximity scans of the whole group. We provide a detailed description of these variables in the supplementary material.

254 Data collection

255 Data collection started on April 28th, 2022, lasted until October 7th, 2023 and
256 included 10 time points. One time point meant running all tasks with all participants.
257 Within each time point, the tasks were organized in three sessions (see Fig. 1), which usually
258 took place on three consecutive days. Session 1 included the population-to-sample and
259 probabilistic-reasoning tasks, session 2 the communicative-cues and ignore-visible-food tasks
260 and session 3 the self-ordered-search and attention-following tasks.

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

266 points.

267 **Analysis, results and discussion**

268 All data files and analysis scripts can be found in the associated online repository
 269 (<https://github.com/ccp-eva/laac2>). The supplementary material provides additional details
 270 on the statistical analysis and the results. In the following, we limit ourselves to describing
 271 only the key analysis and results.

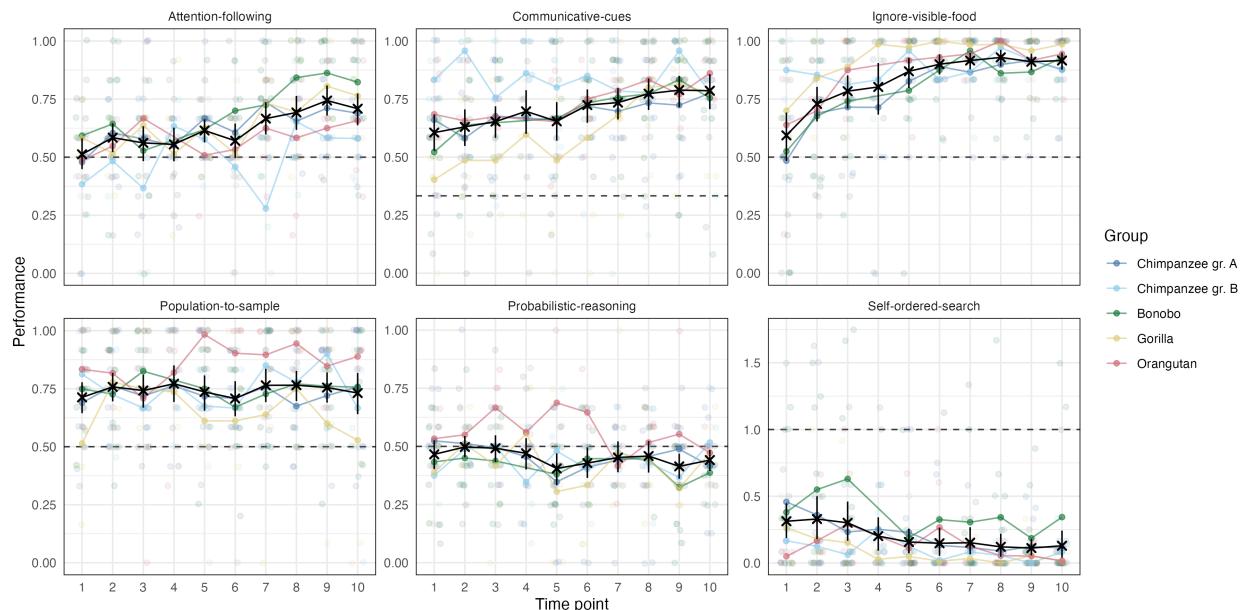


Figure 2. Results from the six cognitive tasks across time points. Black crosses show mean performance at each time point across species (with 95% CI). The sample size varied between time points and can be found in Supplementary Figure 1. Colored dots show mean performance by species. Dashed line shows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

272 To get an overview of the results, we first visualized the data (Fig. 2). Group-level
 273 performance was consistently above chance in the communicative-cues, ignore-visible-food
 274 and population-to-sample tasks. For attention-following, this was the case only from time
 275 point 7 onward and for probabilistic-reasoning, performance was, if anything, below chance.
 276 For the self-ordered-search task, performance was below chance but here lower values reflect

277 better performance (i.e. systematic avoidance of the visible food item). For
278 attention-following, ignore-visible-food, communicative-cues and self-ordered-search there
279 was a steady improvement in performance over time.

280 In the following, we link performance in the tasks across time points to latent variables
281 representing cognitive abilities. We first ask how stable these abilities are over time and how
282 reliably they are measured. Next, we study the correlations between different abilities to
283 explore the internal structure of great ape cognition. Finally, we link performance in the
284 tasks to external predictors to shed light on the sources of individual differences in abilities.
285 Each section uses different statistical techniques which we describe in the respective section.

286 **Stability and reliability**

287 We first asked how stable performance was on a task-level, how stable individual
288 differences were and how reliable the measures were. We used *Structural Equation Modeling*
289 (SEM) [50,51] to address these questions¹. For each task we fit two types of models that
290 addressed different questions. We provide a detailed, mathematical description of the models
291 in the supplementary material.

292 We started with a latent state (LS) model. The goal of this model is to estimate a
293 measurement-error free latent state, representing an individual's cognitive ability, for each
294 time point. We divided the trials from one time point into two test-halves. Roughly
295 speaking, the correlation between these two test-halves is an indicator of measurement
296 precision and used to estimate measurement error (and reliability). Mean changes in
297 task-level performance can be assessed by comparing the means of latent states across
298 subjects for the different time points. Stability of individual differences can be assessed by

¹ SEMs usually use larger sample sizes than available in the present study. [6] reported a simulation study showing that parameters could be accurately estimated using Bayesian estimation techniques and reasonable model restrictions with sample sizes comparable to the one we have here. We lay out the restrictive assumptions we imposed on the parameters in the supplementary material.

299 correlating latent states across different time points.

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

307 Correlations between latent states illustrated varying degrees of stability of individual
308 differences across tasks (Fig. 3B). Attention-following displayed low-to-moderate correlations
309 at early time points (before time point 7), increasing substantially thereafter.

310 Communicative-cues, ignore-visible-food, and self-ordered-search generally showed high
311 correlations between latent states (with time point 1 of ignore-visible-food being an
312 exception). Population-to-sample correlations were consistently high, while
313 probabilistic-reasoning showed generally low, sometimes even negative, correlations,
314 suggesting no stability across time points.

315 Next, we fit a latent state-trait (LST) models. In comparison to the LS models, these
316 models assume that there is a single latent trait, representing an individual's stable cognitive
317 ability, that is the same across time points. This way we can partition variation in
318 performance on a given time point into variance due to the trait (consistency), variance due
319 to the occasion (occasion specificity; 1 - consistency), and measurement error (used to
320 estimate reliability). Like the latent states in the LS model, the trait in the LST model is
321 assumed to be measurement error free [52–54]. Classic LST models assume that the absolute
322 trait values do not change over time. After inspecting the data, we decided to relax this
323 assumption to account for the mean change in performance over time. Thus, we fit LST
324 models that allowed the absolute trait values to change over time. Change over time,

325 however, is seen as change that is the same for all individuals. Stability of individual
 326 differences is reflected in the proportion of variance explained by the trait (consistency).

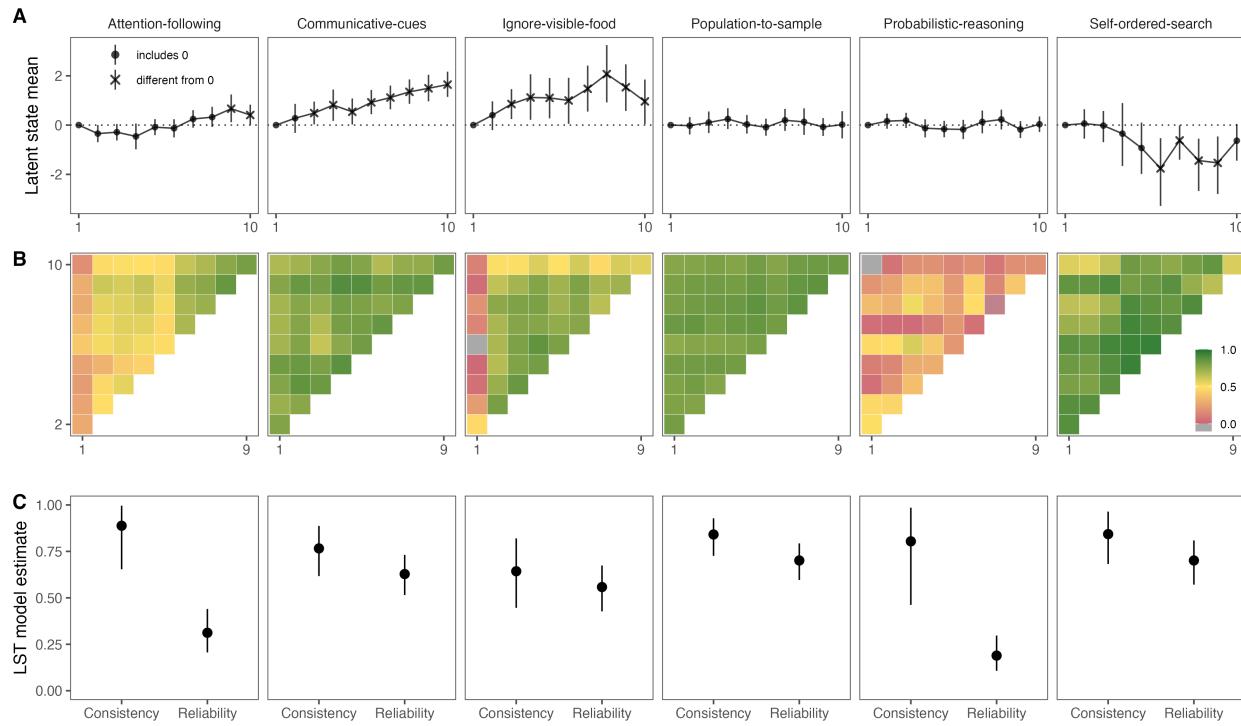


Figure 3. A) Latent mean estimates for each time point by task based on latent state model. Means at time point 1 are set to zero. Shape denotes whether the 95% CrI included zero (dashed line). The sample size varied between time points and can be found in Supplementary Fig. 1. B) Correlations between 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.

327 Consistency estimates varied across tasks (Fig. 3C). In attention-following, the
 328 consistency coefficient was estimated to be 0.89 (95% CI: 0.65 - 0.996), suggesting that
 329 almost 90% of true inter-individual differences were attributable to stable traits. However,
 330 given the low reliability of measurement (see below), this result should be interpreted with

caution because the variability in responses largely reflects measurement error and only to a small extent stable differences between individuals. For communicative-cues, consistency was estimated to be 0.77 (95% CI: 0.62 - 0.89). That is, 77% of the variance was explained by trait differences between individuals. For ignore-visible-food, this number was at 64% (0.64; 95% CI: 0.45 - 0.82). Probabilistic-reasoning showed a similar pattern to attention-following: Consistency was estimated to be high (0.80; 95% CI: 0.46 - 0.98) but reliability was low so that the same restrictions for interpretation apply. Self-ordered-search and population-to-sample had high consistency estimates: 0.70 (95% CI: 0.57 - 0.81) for self-ordered-search and 0.84 (95% CI: 0.73 - 0.93) for population-to-sample.

Reliability of measurement also varied significantly across tasks, based on the LST models (Fig. 3C). For attention-following, reliability was initially low (0.31; 95% CI: 0.21 - 0.44), but was substantially higher when only considering time points 7 and onward (0.66; 95% CI: 0.52 - 0.79). Communicative-cues showed moderate reliability (0.63; 95% CI: 0.52 - 0.73). Ignore-visible-food also had moderate reliability (0.56; 95% CI: 0.43 - 0.67). As mentioned above, probabilistic-reasoning exhibited very low reliability (0.19; 95% CI: 0.11 - 0.30). Population-to-sample showed acceptable reliability (0.70; 95% CI: 0.60 - 0.79). Self-ordered-search also exhibited acceptable reliability levels (0.70; 95% CI: 0.57 - 0.81).

To summarize the SEM results, we saw that the six tasks differed substantially in what they revealed about group- and individual-level variation. What stands out is the widespread change in performance over time. For all tasks except population-to-sample and probabilistic-reasoning we observed an improvement in performance over time. This group-level change, however, has different individual-level interpretations for the different tasks. For communicative-cues, ignore-visible-food and self-ordered-search, individual differences remained relatively stable despite the group-level change suggesting stable individual differences combined with a systematic learning effect across individuals. In contrast, for attention-following, there was little stability in individual differences at earlier

357 time points and only towards the end emerged a more stable ordering of individuals. In
358 combination with the low reliability at earlier time points, this suggests that at least some
359 individuals changed their response strategy in the course of the study. The combination of
360 low reliability, chance-level performance and low correlation of latent states for
361 probabilistic-reasoning suggests that this task is not suited to assess individual differences in
362 probabilistic reasoning abilities in great apes.

363 It is also noteworthy that – across tasks – the reliability estimates are on average lower
364 compared to a previous study testing the same individuals on different tasks [6]. One
365 explanation might be the increase in performance over time, which was not observed by [6].
366 At the beginning of the study, more individuals might have chosen randomly instead of using
367 the available information provided in the task setup and the demonstrations. By definition,
368 random variation is not reliable. With time, more and more individuals started using the
369 available information so that inter-individual differences in how good they are in using it
370 could be detected.

371 Structure

372 To explore the structure of great ape cognition we correlated latent trait estimates for
373 each task. In contrast to raw performance scores, these estimates take into account the
374 reliability of measurement and are considered to be measurement-error free. [6] tested the
375 same individuals and we therefore also include the data from tasks reported there (data from
376 phase 2). Even though the data in the two studies was collected at different time points, we
377 think it is justifiable to analyse them jointly because it is unlikely that changes in cognitive
378 abilities (over and above task-specific training effects that apply to all individuals) occur in
379 this time span. We saved 50 plausible values for the latent trait variables per individual and
380 task after MCMC convergence [see 55], which were combined across tasks and analysed as
381 multiple imputations, obtaining a pooled estimate per correlation with a respective standard
382 error based on the pooling method for multiple imputations suggested by [56].

Figure 4 shows the correlations between trait estimates for the different tasks. Overall,

most correlations were not significantly different from zero (i.e. the 95% CI did include zero).

Because of this low average level of correlations, we decided not to explore models with

higher-order factors and will only interpret specific qualitative patterns.

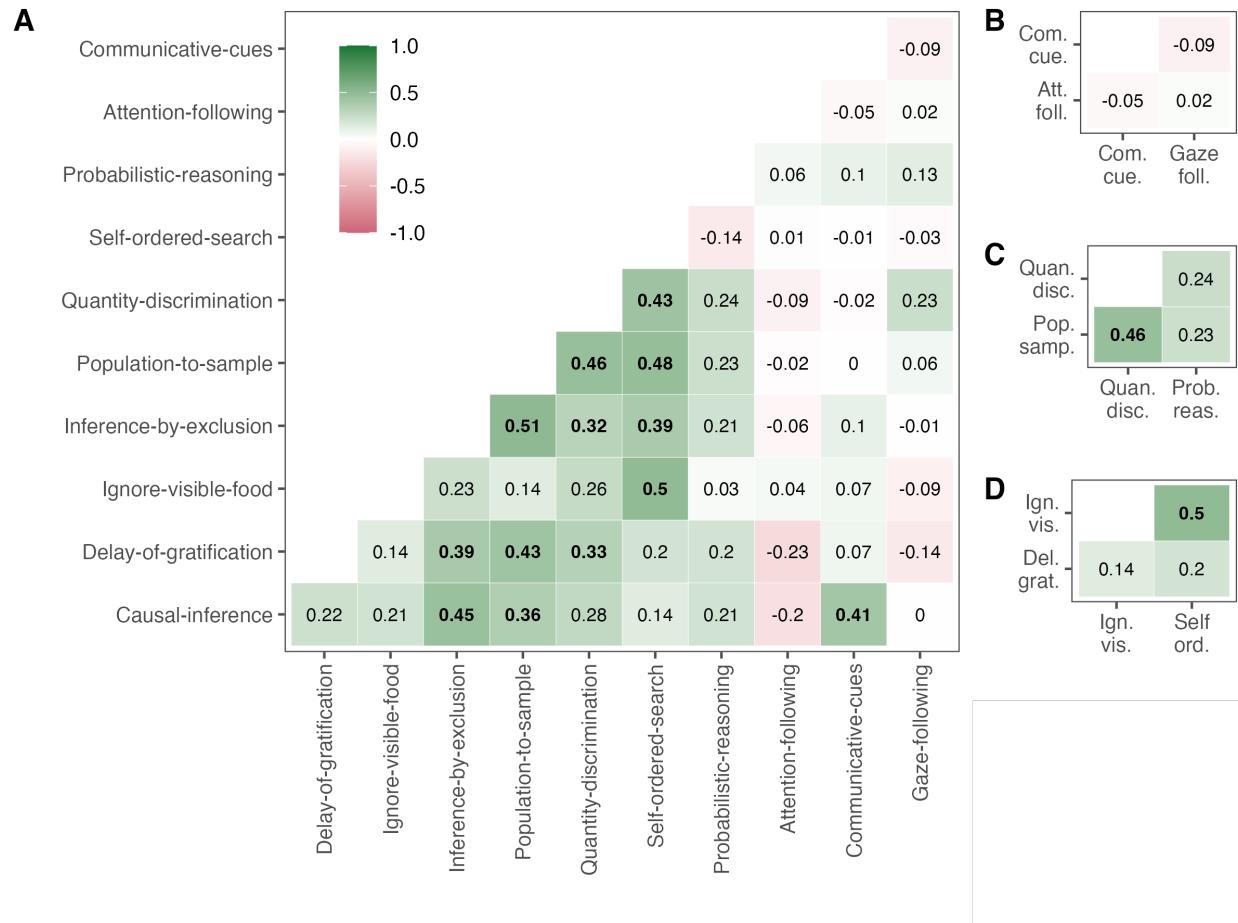


Figure 4. Correlations between trait estimates. Bold correlations have 95% CrI not overlapping with zero. Panels show correlations between A) all, B) social cognition, C) reasoning about quantities, and D) executive functions tasks. The correlation between the two inferential reasoning tasks is not shown in a separate panel but can be found in A). Correlations involving self-ordered-search (coded as number of errors) have been multiplied by -1 so that higher values can be interpreted as better traits for all tasks.

Conceptually, the tasks can be clustered in the following broader domains: *social*

388 *cognition* (attention-following, gaze-following, communicative-cues), *reasoning about*
389 *quantities* (quantity-discrimination, population-to-sample, probabilistic-reasoning), *executive*
390 *functions* (delay-of-gratification, self-ordered-search, ignore-visible-food) and *inferential*
391 *reasoning* (causal-inference, inference-by-exclusion). As a first step, we will evaluate whether
392 we find evidence for such a clustering in the data.

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

394 Furthermore, attention-following and gaze-following did not correlate significantly with any
395 of the other tasks and communicative-cues correlated only with causal-inference – a result we
396 will discuss below. Thus, and in line with previous work [4], we found no evidence for shared
397 cognitive processes in tasks measuring different aspects of social cognition.

398 For the three tasks measuring reasoning about quantities only quantity-discrimination
399 and population-to-sample did correlate significantly. Both tasks require discriminating
400 between different quantities, directly in the case of quantity-discrimination and as part of the
401 decision making process in the case of population-to-sample. Deciding between the samples
402 from the two populations requires discriminating between the relative quantities within each
403 bucket from which the samples were drawn. Probabilistic-reasoning did not correlate with
404 either of the other two quantity tasks (neither did it with any other task). This is not
405 surprising given the results reported above: the observed variation in the probabilistic
406 reasoning task was largely noise and did not reflect systematic individual differences.

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

414 differences in self control and less due to differences in inhibition.

415 Finally, we found a correlation between the two inferential reasoning measures,
416 inference-by-exclusion and causal-inference. This correlation is most likely due to the fact
417 that both tasks involve making inferences about the location of food based on reasoning
418 about its physical properties.

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

426 Finally, when zooming out a bit, a notable cluster including all non-social tasks that
427 reliably measured individual differences (i.e. excluding probabilistic reasoning) emerges. Out
428 of 21 correlations, 12 were significant. All others were positive and numerically close to the
429 significant ones. On a generous view, one might further consider that self-ordered-search and
430 ignore-visible-food had limited variation due to ceiling effects which might have led to an
431 underestimation of the correlations involving these tasks (6 out of 9 non-significant
432 correlations). In sum, one might therefore speculate about commonalities between all
433 non-social tasks. What could these be? We do not know, however, we doubt that it would
434 be a single process shared by all the tasks. More likely is a set of processes that are shared
435 among some of the tasks. We think the best way to find out would be to adopt a
436 process-level perspective on all tasks and build computational cognitive models that
437 explicate the processes involved. This approach could be constrained using the data reported
438 here and, more importantly, it could lead to predictions about which, yet to be designed,
439 tasks should correlate because they share a common set of processes.

440 Predictability

441 In this section, we analysed which external variables accounted for inter- and
442 intra-individual differences in task performance. That is, we asked which of the predictor
443 variables described above predicted performance in the different tasks. Given the large
444 number of predictor variables (14), this question translates to a variable selection problem:
445 selecting a subset of variables from a larger pool. We used the projection predictive inference
446 [57] approach because it is a state-of-the-art procedure that provides an excellent trade-off
447 between model complexity and accuracy [58,59]. The projection prediction approach is a
448 two-step process: The first step consists of building the best predictive model possible, called
449 the reference model. In our case, the reference model is a Bayesian multilevel regression
450 model – fit via `brms` [60] – including all available predictors [61]. In the second step, the goal
451 is to replace the posterior distribution of the reference model with a simpler distribution
452 containing fewer predictors compared to the reference model. The importance of a predictor
453 is assessed by inspecting the mean log-predictive density (`elpd`) and root-mean-squared error
454 (`rmse`) of models containing the predictor compared to models that lack it.

455 The output of the procedure is a ranking of the different predictors. That is, for each
456 task, we get a ranking of how important a predictor is for constructing the simpler
457 replacement distribution. In addition, we can make a qualitative assessment of whether or
458 not a predictor is relevant or not. In addition to the global assessment, we also inspected the
459 projected posterior distribution of the predictors classified as relevant to see how they
460 influenced performance. In the supplementary material we provide a detailed description of
461 the procedure including how the different variables were handled and how the importance of
462 each predictor was assessed.

463 In addition to the external predictors, the models also included a random intercept
464 term for subject ((1 | subject) in `brms` notation). This predictor was handled in a special
465 way in that it was always considered last because it would otherwise have soaked up most of

466 the variance before the other predictors would have had a chance to explain any of it.

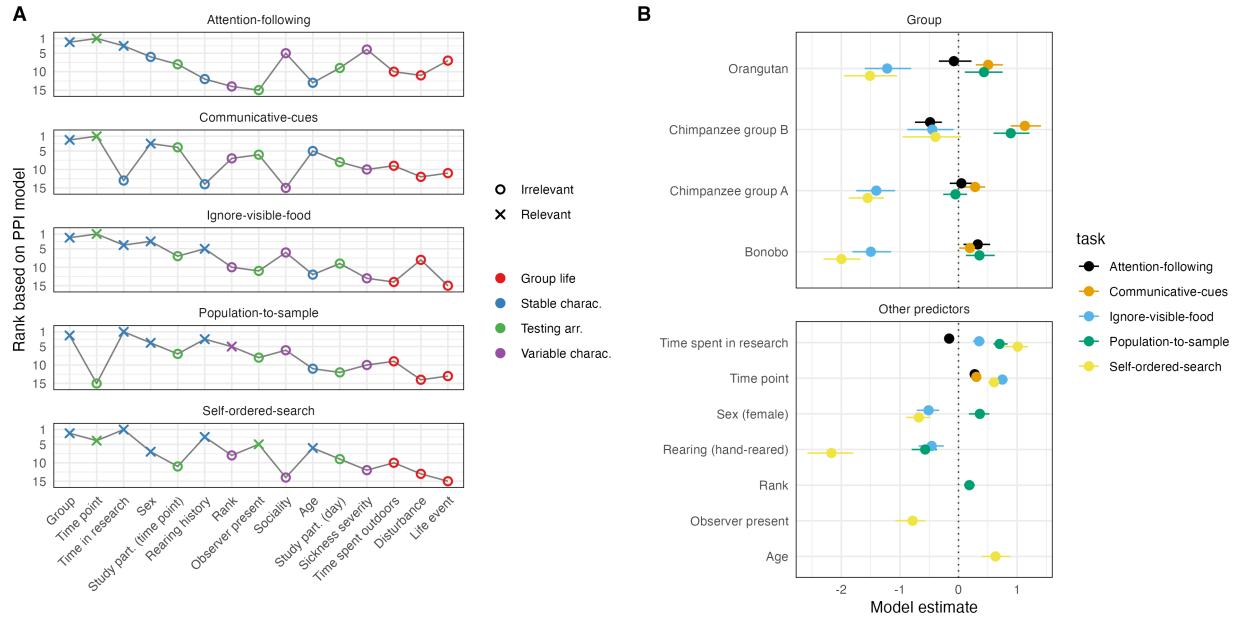


Figure 5. A) Predictor ranking and selection based on PPI models. Crosses mark predictors that were selected to be relevant based on the PPI models. Color shows the broader category each predictor belongs to. The x-axis is sorted by the average rank across tasks. B) Posterior model estimates for the 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).

467 Fig. 5A summarizes the selected predictors across tasks. For all tasks, the random
 468 intercept term improved model fit the most (not shown in Fig. 5A). In line with results
 469 reported by Bohn et al. (2023), this suggests that genetic predispositions and/or
 470 idiosyncratic developmental processes, which operate on time-scales longer than what we
 471 captured in our study, accounted for a substantial portion of the variance in cognitive
 472 abilities between individuals.

473 However, for two tasks, other predictors had a comparable explanatory power –
 474 something that was not observed in [6]. For population-to-sample, **time spent in**
 475 **research** improved the model fit even more than adding the random intercept at the end

476 did. This could be interpreted that performance in this task strongly depends on having
477 learned to pay attention to stimuli and the human experimenter. For ignore-visible-food,
478 **time point** had an influence exceeding that of the random intercept term. We think this
479 result reflects the strong within-task learning effect across subjects. Because performance
480 increased substantially with time, most of the variation captured by **time point** exceeded
481 the variation between individuals.

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

489 The predictor selected most often was **group**. It was the only predictor that was
490 selected as relevant for all tasks. However, differences between groups were variable in that
491 the ranking of the groups changed from task to task (Fig. 5B). For example, gorillas
492 performed best in ignore-visible-food and self-ordered-search, the chimpanzee group B
493 performed best in communicative-cues and population-to-sample and the bonobos performed
494 best in attention-following. This speaks against clear species or group differences in general
495 cognitive performance. Again, the most likely explanation for group differences is an
496 interaction between species-specific dispositions and individual- / task-level developmental
497 processes.

498 The predictors that were selected more than once influenced performance in variable
499 ways (Fig. 5B). As mentioned above, **time point** always had a positive effect because
500 performance increased with time. Whenever **rearing** was selected to be relevant,
501 mother-reared individuals outperformed others. Time spent in research had a positive

502 effect, suggesting that more experience with research [or researchers, see 62] leads to better
503 performance, except for attention-following. The effect of **sex** was variable in that females
504 outperformed males in population-to-sample but males outperformed females in
505 self-ordered-search and ignore-visible-food.

506 **General Discussion**

507 In the present study, we investigated the stability, structure and predictability of great
508 ape cognition across a broad range of domains, including social cognition, reasoning about
509 quantities, executive functions, and inferential reasoning. We repeatedly administered six
510 tasks to a comparatively large sample of great apes a total of 10 times over a period of 1.5
511 years. Group-level results varied by task: while some tasks demonstrated substantial changes
512 over time, others remained relatively stable. The tasks also differed in how reliably they
513 measured individual differences, ranging from very poor (probabilistic-reasoning) to very
514 good (population-to-sample, self-ordered-search). A significant portion of the observed
515 variance in performance could be attributed to stable differences in cognitive abilities
516 between individuals. However, these individual differences were not strongly associated
517 across all tasks; instead most non-social tasks were correlated while social tasks correlated
518 neither with each other nor with other tasks.. Finally, individual differences in cognitive
519 abilities were better predicted by stable, individual-specific characteristics compared to
520 transient aspects of everyday experience.

521 The observed substantial changes in performance over time highlight the plasticity of
522 cognition in great apes. Even though individual differences were stable – indicating that
523 individuals improved at similar rates – our findings show that adult apes, including older
524 individuals, are capable of learning and cognitive improvement. A case in point is the
525 chimpanzee B group, which consisted exclusively of adults, some of whom were in their 50s.
526 This contrasts with earlier work which suggested a decline in cognitive performance, in
527 particular executive functions, with age [63–65]. However, earlier findings might have been

528 driven by cohort effects in that longitudinal decline within individuals was substantially
529 smaller compared to cross-section differences between age groups [66]. In any case, this
530 underscores the importance of longitudinal studies to study the dynamics of cognitive
531 development, not just early but also late in life.

532 The tasks varied substantially in their quality of measurement. This finding
533 emphasizes the importance of rigorously assessing measurement properties before including
534 tasks in cognitive test batteries or collecting data from large samples with the goal of
535 assessing individual differences [see also 67,68]. The reliability of measurement has profound
536 implications for the conclusions that can be drawn [69,70]. For instance, the
537 probabilistic-reasoning task showed no meaningful correlations with other tasks, which might
538 suggest that probabilistic reasoning is an isolated cognitive ability. However, the lack of
539 correlation—paired with chance-level performance—was more likely due to the task failing to
540 measure anything reliably, with variation in performance being predominantly noise. The
541 communicative-cues task, on the other hand, reliably measured individual differences but did
542 not correlate with any of the other tasks, suggesting that it does not share cognitive
543 processes with them.

544 We found no evidence for a *g*-factor explaining much of the differences between
545 individuals [contra 15]. Compared to work with human participants and also to earlier ape
546 studies [4,15,18], the sample we tested could be considered small. However, we collected a
547 large number of data points for each individual, and our analytical approach explicitly
548 accounted for measurement reliability. Thus, we believe the lack of strong correlations across
549 tasks reflects a genuine finding rather than noise. This pattern also aligns with previous
550 work and animal cognition research more broadly [71]. For example, when conducting a
551 confirmatory factor analysis on their data, [4] found that less than half of the tasks in the
552 PCTB loaded on any of the theoretically proposed factors, and only 10 of the 105 bivariate
553 correlations were significant. [18] found that only three out of 36 bivariate correlations

554 between executive functions tasks were significantly different from zero. Moving forward, we
555 propose to move away from a domain-level perspective to a process-level perspective. That
556 is, instead of classifying tasks based on their domain of application (e.g., reasoning about the
557 physical or social world), one should identify the cognitive processes involved in a task and
558 generate predictions about correlations between tasks based on process-level commonalities.
559 Case in point is the correlation observed between causal-inference and communicative-cues,
560 which can only be explained by a process-level perspective.

561 Finally, this study, alongside findings from [6], highlights that the origins of individual
562 differences in great ape cognitive abilities most likely lie deeply embedded in the ontogenetic
563 – and perhaps genetic – history of individuals. Efforts to explain these differences by using
564 easily measurable variables, such as age, sex, or rank, proved unproductive. Of these, only
565 group emerged as a relevant predictor across all tasks. Notably, in this study, group is not
566 synonymous with species: the two chimpanzee groups differed substantially across tasks.
567 This underscores the importance of studying within-species variation, rather than focusing
568 solely on between-species differences. On its own, the group variable has limited explanatory
569 power because it encapsulates a variety of factors, including age, social dynamics and genetic
570 differences. Altogether, these findings highlight the need for longitudinal studies that begin
571 as early in life as possible to truly understand the developmental roots of individual
572 differences in great ape cognition.

573

References

- 574 1. Völter CJ, Tinklenberg B, Call J, Seed AM. 2018 Comparative psychometrics: Establishing what differs is central to understanding what evolves. *Philosophical Transactions of the Royal Society B: Biological Sciences* **373**, 20170283.
- 575 2. Shaw RC, Schmelz M. 2017 Cognitive test batteries in animal cognition research: Evaluating the past, present and future of comparative psychometrics. *Animal Cognition* **20**, 1003–1018.
- 576 3. Thornton A, Lukas D. 2012 Individual variation in cognitive performance: Developmental and evolutionary perspectives. *Philosophical Transactions of the Royal Society B: Biological Sciences* **367**, 2773–2783.
- 577 4. Herrmann E, Hernández-Lloreda MV, Call J, Hare B, Tomasello M. 2010 The structure of individual differences in the cognitive abilities of children and chimpanzees. *Psychological Science* **21**, 102–110.
- 578 5. Watson SK *et al.* 2018 Chimpanzees demonstrate individual differences in social information use. *Animal Cognition* **21**, 639–650.
- 579 6. Bohn M, Eckert J, Hanus D, Lugauer B, Holtmann J, Haun DB. 2023 Great ape cognition is structured by stable cognitive abilities and predicted by developmental conditions. *Nature Ecology & Evolution* **7**, 927–938.
- 580 7. Fröhlich M, Wittig RM, Pika S. 2019 The ontogeny of intentional communication in chimpanzees in the wild. *Developmental science* **22**, e12716.

- 581 8. Berdugo S, Cohen E, Davis A, Matsuzawa T, Carvalho S. 2023 Stable long-term
individual variation in chimpanzee technological efficiency. *bioRxiv*, 2023–11.
- 582 9. Deaner RO, Van Schaik CP, Johnson V. 2006 Do some taxa have better domain-general
cognition than others? A meta-analysis of nonhuman primate studies. *Evolutionary
Psychology* **4**, 147470490600400114.
- 583 10. Burkart JM, Schubiger MN, Schaik CP van. 2017 The evolution of general intelligence.
Behavioral and Brain Sciences **40**.
- 584 11. Reader SM, Hager Y, Laland KN. 2011 The evolution of primate general and cultural
intelligence. *Philosophical Transactions of the Royal Society B: Biological Sciences*
366, 1017–1027.
- 585 12. Herrmann E, Call J, Hernández-Lloreda MV, Hare B, Tomasello M. 2007 Humans
have evolved specialized skills of social cognition: The cultural intelligence hypothesis.
Science **317**, 1360–1366.
- 586 13. Fichtel C, Dinter K, Kappeler PM. 2020 The lemur baseline: How lemurs compare to
monkeys and apes in the primate cognition test battery. *PeerJ* **8**, e10025.
- 587 14. Schmitt V, Pankau B, Fischer J. 2012 Old world monkeys compare to apes in the
primate cognition test battery. *PloS one* **7**, e32024.
- 588 15. Hopkins WD, Russell JL, Schaeffer J. 2014 Chimpanzee intelligence is heritable.
Current Biology **24**, 1649–1652.

- 589 16. Beran MJ, Hopkins WD. 2018 Self-control in chimpanzees relates to general intelligence.
Current Biology **28**, 574–579.
- 590 17. Kaufman AB, Reynolds MR, Kaufman AS. 2019 The structure of ape (hominoidea) intelligence. *Journal of Comparative Psychology* **133**, 92.
- 591 18. Völter CJ *et al.* 2022 The structure of executive functions in preschool children and chimpanzees. *Scientific Reports* **12**, 1–16.
- 592 19. Dunbar R, Shultz S. 2017 Why are there so many explanations for primate brain evolution? *Philosophical Transactions of the Royal Society B: Biological Sciences* **372**, 20160244.
- 593 20. Rosati AG. 2017 Foraging cognition: Reviving the ecological intelligence hypothesis. *Trends in cognitive sciences* **21**, 691–702.
- 594 21. Henke-von der Malsburg J, Kappeler PM, Fichtel C. 2020 Linking ecology and cognition: Does ecological specialisation predict cognitive test performance? *Behavioral Ecology and Sociobiology* **74**, 154.
- 595 22. Rosati AG, Santos LR. 2017 Tolerant barbary macaques maintain juvenile levels of social attention in old age, but despotic rhesus macaques do not. *Animal behaviour* **130**, 199–207.

- 596 23. Joly M, Micheletta J, De Marco A, Langermans JA, Sterck EH, Waller BM. 2017
Comparing physical and social cognitive skills in macaque species with different
degrees of social tolerance. *Proceedings of the Royal Society B: Biological Sciences*
284, 20162738.
- 597 24. Amici F, Aureli F, Call J. 2008 Fission-fusion dynamics, behavioral flexibility, and
inhibitory control in primates. *Current Biology* **18**, 1415–1419.
- 598 25. Kaigaishi Y, Nakamichi M, Yamada K. 2019 High but not low tolerance populations
of Japanese macaques solve a novel cooperative task. *Primates* **60**, 421–430.
- 599 26. Piantadosi ST, Kidd C. 2016 Extraordinary intelligence and the care of infants.
Proceedings of the National Academy of Sciences **113**, 6874–6879.
- 600 27. Schubiger MN, Fichtel C, Burkart JM. 2020 Validity of cognitive tests for non-human
animals: Pitfalls and prospects. *Frontiers in Psychology* **11**, 1835.
- 601 28. ManyPrimates *et al.* 2022 The evolution of primate short-term memory. *Animal
Behavior and Cognition* **9**, 428–516.
- 602 29. ManyPrimates *et al.* 2019 Establishing an infrastructure for collaboration in primate
cognition research. *PLoS One* **14**, e0223675.
- 603 30. Sih A, Sinn DL, Patricelli GL. 2019 On the importance of individual differences in
behavioural skill. *Animal Behaviour* **155**, 307–317.

- 604 31. Berio L, Moore R. 2023 Great ape enculturation studies: A neglected resource in cognitive development research. *Biology & Philosophy* **38**, 17.
- 605 32. Bard KA, Bakeman R, Boysen ST, Leavens DA. 2014 Emotional engagements predict and enhance social cognition in young chimpanzees. *Developmental Science* **17**, 682–696.
- 606 33. Van Leeuwen EJ, DeTroy SE, Kaufhold SP, Dubois C, Schütte S, Call J, Haun DB. 2021 Chimpanzees behave prosocially in a group-specific manner. *Science advances* **7**, eabc7982.
- 607 34. Altschul DM, Wallace EK, Sonnweber R, Tomonaga M, Weiss A. 2017 Chimpanzee intellect: Personality, performance and motivation with touchscreen tasks. *Royal Society Open Science* **4**, 170169.
- 608 35. Hopper LM, Price SA, Freeman HD, Lambeth SP, Schapiro SJ, Kendal RL. 2014 Influence of personality, age, sex, and estrous state on chimpanzee problem-solving success. *Animal cognition* **17**, 835–847.
- 609 36. Brosnan SF, Hopper LM, Richey S, Freeman HD, Talbot CF, Gosling SD, Lambeth SP, Schapiro SJ. 2015 Personality influences responses to inequity and contrast in chimpanzees. *Animal behaviour* **101**, 75–87.
- 610 37. Carter AJ, Marshall HH, Heinsohn R, Cowlishaw G. 2014 Personality predicts the propensity for social learning in a wild primate. *PeerJ* **2**, e283.

- 611 38. Matzel LD, Sauce B. 2017 Individual differences: Case studies of rodent and primate
intelligence. *Journal of Experimental Psychology: Animal Learning and Cognition* **43**,
325.
- 612 39. Boogert NJ, Madden JR, Morand-Ferron J, Thornton A. 2018 Measuring and under-
standing individual differences in cognition. *Philosophical Transactions of the Royal
Society B: Biological Sciences* **373**, 20170280.
- 613 40. Griffin AS, Guillette LM, Healy SD. 2015 Cognition and personality: An analysis of
an emerging field. *Trends in Ecology & Evolution* **30**, 207–214.
- 614 41. Kaminski J, Call J, Tomasello M. 2004 Body orientation and face orientation: Two
factors controlling apes' begging behavior from humans. *Animal cognition* **7**, 216–223.
- 615 42. 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**, e0187451.
- 616 43. Völter CJ, Tinklenberg B, Call J, Seed AM. 2022 Inhibitory control and cue relevance
modulate chimpanzees' (*pan troglodytes*) performance in a spatial foraging task. *Jour-
nal of Comparative Psychology* **136**, 105.
- 617 44. Hanus D, Call J. 2014 When maths trumps logic: Probabilistic judgements in chim-
panzees. *Biology letters* **10**, 20140892.
- 618 45. Rakoczy H, Clüver A, Saucke L, Stoffregen N, Gräbener A, Migura J, Call J. 2014
Apes are intuitive statisticians. *Cognition* **131**, 60–68.

- 619 46. Eckert J, Call J, Hermes J, Herrmann E, Rakoczy H. 2018 Intuitive statistical inferences
in chimpanzees and humans follow weber's law. *Cognition* **180**, 99–107.
- 620 47. Völter CJ, Mundry R, Call J, Seed AM. 2019 Chimpanzees flexibly update working
memory contents and show susceptibility to distraction in the self-ordered search task.
Proceedings of the Royal Society B **286**, 20190715.
- 621 48. Petrides M. 1995 Impairments on nonspatial self-ordered and externally ordered
working memory tasks after lesions of the mid-dorsal part of the lateral frontal cortex
in the monkey. *Journal of Neuroscience* **15**, 359–375.
- 622 49. Diamond A, Prevor MB, Callender G, Druin DP. 1997 Prefrontal cortex cognitive
deficits in children treated early and continuously for PKU. *Monographs of the society
for research in child development*, i–206.
- 623 50. Bollen KA. 1989 *Structural equations with latent variables*. John Wiley & Sons.
- 624 51. Hoyle RH. 2012 *Handbook of structural equation modeling*. Guilford press.
- 625 52. Steyer R, Ferring D, Schmitt MJ. 1992 States and traits in psychological assessment.
European Journal of Psychological Assessment **8**, 79–98.
- 626 53. Steyer R, Mayer A, Geiser C, Cole DA. 2015 A theory of states and traits—revised.
Annual Review of Clinical Psychology **11**, 71–98.
- 627 54. Geiser C. 2020 *Longitudinal structural equation modeling with mplus: A latent state-
trait perspective*. Guilford Publications.

- 628 55. Asparouhov T, Muthén B. 2010 Bayesian analysis in mplus: Technical implementation.
- 629 56. Rubin DB. 1987 *Multiple imputation for nonresponse in surveys*. New York: John Wiley & Sons.
- 630 57. Piironen J, Paasiniemi M, Vehtari A. 2020 Projective inference in high-dimensional problems: Prediction and feature selection. *Electronic Journal of Statistics* **14**, 2155–2197. (doi:10.1214/20-EJS1711)
- 631 58. Piironen J, Vehtari A. 2017 Comparison of bayesian predictive methods for model selection. *Statistics and Computing* **27**, 711–735. (doi:10.1007/s11222-016-9649-y)
- 632 59. Pavone F, Piironen J, Bürkner P-C, Vehtari A. 2020 Using reference models in variable selection.
- 633 60. Bürkner P-C. 2017 brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software* **80**, 1–28.
- 634 61. Catalina A, Bürkner P-C, Vehtari A. 2020 Projection predictive inference for generalized linear and additive multilevel models. *arXiv preprint arXiv:2010.06994*
- 635 62. Damerius LA *et al.* 2017 Orientation toward humans predicts cognitive performance in orang-utans. *Scientific reports* **7**, 40052.

- 636 63. Lacreuse A, Raz N, Schmidtke D, Hopkins WD, Herndon JG. 2020 Age-related decline
in executive function as a hallmark of cognitive ageing in primates: An overview of
cognitive and neurobiological studies. *Philosophical Transactions of the Royal Society
B* **375**, 20190618.
- 637 64. Lacreuse A, Parr L, Chennareddi L, Herndon JG. 2018 Age-related decline in cognitive
flexibility in female chimpanzees. *Neurobiology of aging* **72**, 83–88.
- 638 65. Manrique HM, Call J. 2015 Age-dependent cognitive inflexibility in great apes. *Animal
Behaviour* **102**, 1–6.
- 639 66. Hopkins WD, Mareno MC, Neal Webb SJ, Schapiro SJ, Raghanti MA, Sherwood CC.
2021 Age-related changes in chimpanzee (*pan troglodytes*) cognition: Cross-sectional
and longitudinal analyses. *American Journal of Primatology* **83**, e23214.
- 640 67. Cauchoux M *et al.* 2018 The repeatability of cognitive performance: A meta-analysis.
Philosophical Transactions of the Royal Society B: Biological Sciences **373**, 20170281.
- 641 68. Soha JA, Peters S, Anderson RC, Searcy WA, Nowicki S. 2019 Performance on tests
of cognitive ability is not repeatable across years in a songbird. *Animal Behaviour*
158, 281–288.
- 642 69. Fried EI, Flake JK. 2018 Measurement matters. *APS Observer* **31**.
- 643 70. Hedge C, Powell G, Sumner P. 2018 The reliability paradox: Why robust cognitive
tasks do not produce reliable individual differences. *Behavior Research Methods* **50**,
1166–1186.

- ⁶⁴⁴ 71. Poirier M-A, Kozlovsky DY, 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**, 20201853.