A baseline for inferences about human cognitive evolution: structure, stability and predictability of great ape cognition

Manuel Bohn[1], Johanna Eckert[1], Daniel Hanus[1], Benedikt Lugauer[2], Jana Holtmann[2,*], & Daniel Haun[1,*]

[1] Department of Comparative Cultural Psychology, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

[2] Psychologische Hochschule Berlin, Berlin, Germany

[*] Shared senior authorship

Author Note

We thank Damilola Olaoba, Anna Wolff and Nico Eisenbrenner for help with the data collection.

The authors made the following contributions. Manuel Bohn: Conceptualization, Analysis, Writing - Original Draft Preparation, Writing - Review & Editing; Johanna Eckert: Conceptualization, Writing - Review & Editing; Daniel Hanus: Conceptualization, Writing - Review & Editing; Benedikt Lugauer: Analysis, Writing - Review & Editing; Jana Holtmann: Analysis, Writing - Review & Editing; Daniel Haun: Conceptualization, Writing - Review & Editing.

Correspondence concerning this article should be addressed to Manuel Bohn, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany. E-mail: manuel_bohn@eva.mpg.de

Abstract

One or two sentences providing a **basic introduction** to the field, comprehensible to a scientist in any discipline.

Two to three sentences of **more detailed background**, comprehensible to scientists in related disciplines.

One sentence clearly stating the **general problem** being addressed by this particular study.

One sentence summarizing the main result (with the words "**here we show**" or their equivalent).

Two or three sentences explaining what the **main result** reveals in direct comparison to what was thought to be the case previously, or how the main result adds to previous knowledge.

One or two sentences to put the results into a more **general context**.

Two or three sentences to provide a **broader perspective**, readily comprehensible to a scientist in any discipline.

Keywords: keywords

Word count: X

A baseline for inferences about human cognitive evolution: structure, stability and predictability of great ape cognition

Introduction

In their quest for understanding the evolution of the human mind, psychologists and cognitive scientists face one major obstacle: cognition does not fossilize. Instead of directly studying the cognitive abilities of our extinct ancestors, we have to rely on backward inferences. We can study fossilized skulls and crania to approximate brain size and structure and use this information to infer cognitive abilities (Coqueugniot, Hublin, Veillon, Houët, & Jacob, 2004; Gunz et al., 2020). We can study the material culture left behind by our ancestors and try to infer its cognitive complexity experimentally (Coolidge & Wynn, 2016; Currie & Killin, 2019; Haslam et al., 2017). Yet, the archaeological record is sparse and only goes back so far in time. Thus, one of the most fruitful approaches to cognitive evolution is the comparative method. By studying extant species of primates, we make backward inferences about the last common ancestor. If species A and B both show cognitive ability X, the last common ancestor of A and B most likely also had ability X (MacLean et al., 2012; Martins & Martins, 1996). To make inferences about the most recent events in human cognitive evolution, we have to study and compare humans and the great apes. Such approach has been highly productive and was the basis for numerous theories about human uniqueness (Heyes, 2018; Laland & Seed, 2021; Penn, Holyoak, & Povinelli, 2008; Tomasello, 2019).

However, using the comparative method in this way requires a strong great ape baseline. That is, it takes a solid and robust way of describing the great ape mind in order to map out how it differs from that of humans. What kind of empirical evidence is required to infer such a baseline? First, group-level results should be stable. Our inferences about the cognitive abilities that great apes – as a group – do or do not have do based on the data we collect today should not change if we repeat the study tomorrow. Second, individual

differences in cognitive abilities should be reliable. That is, methods and procedures should also reliably measure cognitive abilities on an individual level. This is a prerequisite for investigating the relations between different tasks in order to map out the internal structure of great ape cognition (Shaw & Schmelz, 2017; Thornton & Lukas, 2012; Volter, Tinklenberg, Call, & Seed, 2018). Finally, and related to the second point, individual differences should be predictable. Understanding great ape cognition means that we can point to external variables that induce variation in cognitive performance and development. Recently, a number of concerns have been voiced, questioning whether the conventional way of conducting comparative research is suited to provide the empirical basis for inferring such a baseline (Farrar & Ostojic, 2019; ManyPrimates et al., 2019; Schubiger, Fichtel, & Burkart, 2020; Stevens, 2017). A key point in this criticism is that most research simply assumes that the three requirements outlined above are met without testing them empirically. The project reported here directly addresses this fundamental problem.

The prototypical study in comparative research still involves only a handful of individuals from a single species tested with in one cognitive task (see ManyPrimates et al., 2019 for a review). There are, however, several notable exceptions that undertook great effort to provide a more comprehensive picture of the nature and structure of great ape cognition (Beran & Hopkins, 2018; Hopkins, Russell, & Schaeffer, 2014; MacLean et al., 2014; Wobber, Herrmann, Hare, Wrangham, & Tomasello, 2014). Herrmann and colleagues (Herrmann, Call, Hernández-Lloreda, Hare, & Tomasello, 2007) tested a large sample of great apes (chimpanzees and orangutans) and human children in a range of tasks from different cognitive domains. The results indicated pronounced group-level differences between great apes and humans in the social, but not the physical domain. Furthermore, relations between the tasks pointed to a different internal structure of cognition, with a distinct social cognition factor for humans but not great apes (Herrmann, Hernández-Lloreda, Call, Hare, & Tomasello, 2010). Völter and colleagues [] focused on the structure of executive functions. Based on a multi-trait multi-method approach they developed a new test battery to assess

memory updating, inhibition, and attention shifting in chimpanzees and human children. Overall, they found low correlations between tasks and thus no clear support for any of the structures put forward by theoretical models built around adult human data.

Despite their seminal contributions to the field, these studies suffer from the same three shortcomings outlined above. First, it is unclear if the results are stable. That is, if the same individuals were tested again, would we see the same results and arrive at the same conclusions about absolute differences between species. Second, the psychometric properties of the tasks are unknown and it is thus unclear if, for example, low correlations between tasks reflect a genuine lack of shared cognitive processes or simply measurement imprecision. Finally, it remains unclear what causes individual differences – which individual characteristics and experiences predict cognitive performance and development.

The studies reported below seek to solidify the empirical grounds of the great ape baseline. For one-and-a-half years, every two weeks we administered a set of five cognitive tasks (see Figure 1)) to the same population of great apes ($N = 43$). The tasks spanned across cognitive domains (social cognition, causal cognition, numerical reasoning, executive functions) and were based on published procedures widely used in the field of comparative psychology. In addition to the cognitive data, we continuously collected data for more than a dozen variables that capture stable and variable aspects of our participants' life and used this to predict inter- and intra-individual variation cognitive performance. Data collection was split into two phases. After Phase 1 (14 data collection time points), we analysed the data and registered the results (https://osf.io/7qyd8). Phase 2 lasted for another 14 time points and served to replicate and extend Phase 1. this approach allowed us to test a) how stable group level results are, b) how reliable individual differences are, c) how individual differences are structured and d) what predicts cognitive performance.

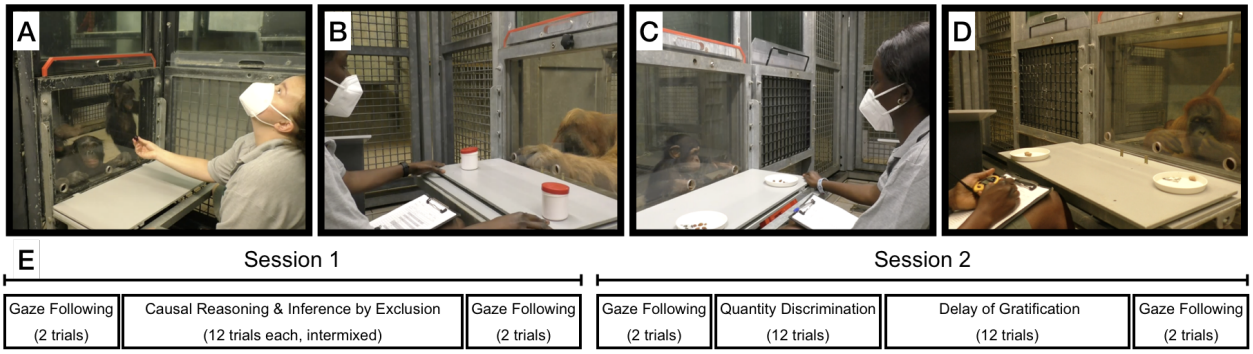

Figure 1. Setup used for the six tasks. A) Gaze following: the experimenter looked to the ceiling. We coded if the ape followed gaze. B) Causal reasoning: food was hidden in one of two cup, the baited cup was shaken (food produced a sound) and apes had to choose the shaken cup to get food. Inference by exclusion: food was hidden in one of two cups. The empty cup was shaken (no sound) so apes had to choose the non-shaken cup to get food. C) Quantity discrimination: Small pieces of food were presented on two plates (5 vs. 7 items); we coded if subjects chose the larger amount. D) Delay of gratification (only Phase 2): to receive a larger reward, the subject had to wait and forgo a smaller, immediately accesible, reward. E) Order of task presentation and trial numbers

Results

Stability of group-level performance

Group-level performance was largely stable or followed clear temporal patterns (see Figure 2). The causal inference and quantity discrimination tasks were the most robust: in both cases performance was clearly different from chance across both phases with no apparent change over time. The rate of gaze following declined in the beginning of Phase 1 but then settled on a low but stable level until the end of Phase 2. This pattern was expected given that following the experimenters gaze was never rewarded – neither explicitly with food or by bringing something interesting to the subject's attention. The inference by exclusion task showed an inverse pattern with group-level performance being at chance-level

for most of Phase 1, followed by a small, but steady, increase throughout Phase 2. These temporal patterns most likely reflect training (or habituation) effects that are a *consequence* of the repeated testing. Performance in the delay of gratification task (Phase 2 only) was slightly variable, but within the same general range. In sum, performance was very robust in that time points generally licensed the same group-level conclusions. The tasks appeared well suited to study group-level performance.

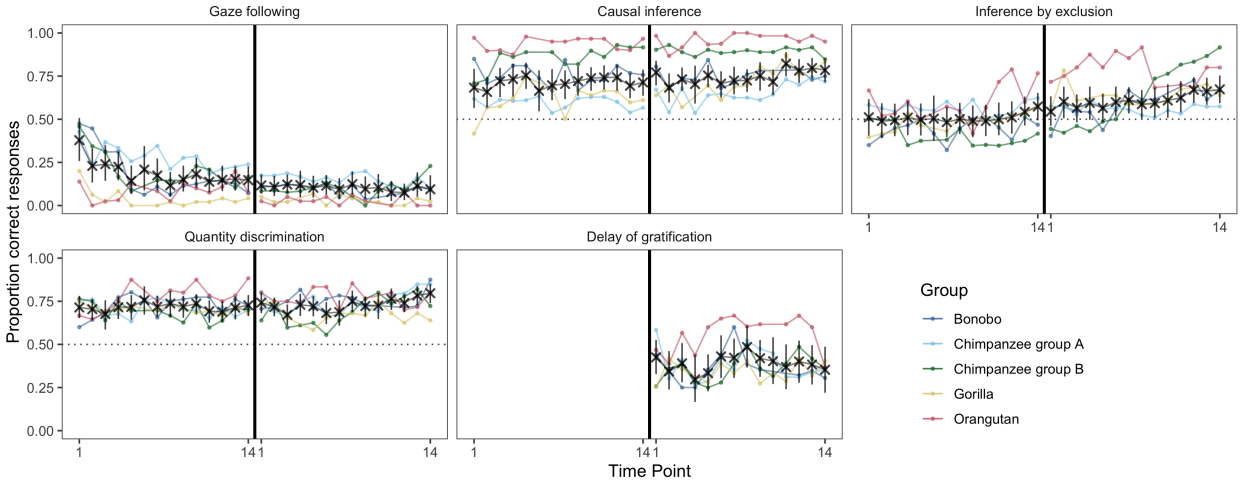

Figure 2. Results from the five cognitive tasks across time points. Black crosses show mean performance at each time point across species (with 95% CI). Colored dots show mean performance by species. Dashed line shows the chance level whenever applicable. The vertical back line marks the transition between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

Reliability of individual differences

Stable group-level performance does not imply stable individual differences. In fact, a well-known paradox in human psychology states that some of the most robust – on a group level – cognitive tasks do not produce reliable individual differences (Hedge, Powell, & Sumner, 2018). In a second step, we therefore assessed the reliability of our five tasks. For that, we correlated the performance at the different time points in each task. Figure 3 visualizes these raw re-test correlations. Correlations were generally high – exceptionally

high for animal cognition standards (Cauchoix et al., 2018) – with higher values for time points closer together (Uher, 2011). The quantity discrimination was less reliable compared to the other tasks.

What stands out in this is that *stability does not imply reliability* - and vice versa. The quantity discrimination task showed robust group-level performance above chance but relatively poor re-test reliability. Group-level performance in the inference by exclusion and gaze following tasks changed over time but was highly reliable on an individual level. Taken together, the majority of tasks was well suited to study individual differences.

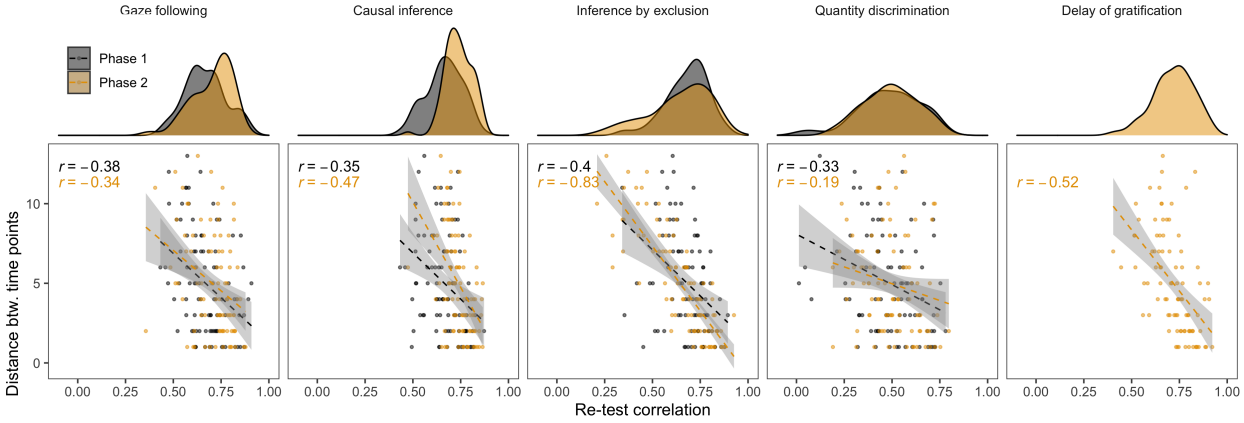

Figure 3. Top: Distribution of re-test correlation coefficients between time points for each task. Bottom: Correlations between re-test reliability coefficients and temporal distance between the testing time points.

Structure of individual differences

Next, we investigated the structure of these individual differences. For this, we used structural equation modelling, in particular latent state-trait models with autoregressive effects (LST-AR) (Eid, Holtmann, Santangelo, & Ebner-Priemer, 2017). These models estimate latent variables which are assumed to be measurement-error free because they are estimated taking into account the reliability of the task. In the LSTM-AR context, reliability is the correlation between (measurement specific) test-halves.

We used these models to partition the inter-individual variance in performance in the five tasks into three different components. First, variance explained by latent traits (*consistency*). In the present context, one can think of a latent trait as a stable cognitive ability (e.g. ability to make causal inferences). Second, variance explained by latent states (*Occasion specificity*), which can be interpreted as time-specific, variable psychological conditions (e.g. variations in performance due to being attentive or inattentive). Finally, variance explained by autoregressive effects (*Predictability*). These effects capture the idea that there are temporal dependencies between time points in that the variance not explained by the trait at time point t can be used to predict performance at $t + 1$. Here, one can think of them as systematic developmental effects that may arise given the longitudinal nature of our data.

Individual differences were largely explained by stable differences in cognitive abilities. Across tasks and phases, around 70% of variance was accounted for by latent trait differences and around 30% by state differences and autoregressive effects (Figure 4A). The high reliability estimates show that these latent variables accounted for most of the variance in raw test scores – with the quantity discrimination task being, once again, an exception.

There are, however, interesting exceptions to this overall pattern. Most notably, for the inference by exclusion and quantity discrimination tasks, predictability was very high in Phase 2 (Figure 4A). This suggests that *some* individuals showed a marked change in their performance over time. These individuals started to deviate from what the model predicted for them based on the latent trait and then continuously developed in a certain direction. If all individuals would have changed in the same way, this would have been accounted for by the latent trait. Looking at Figure 2, this suggests that the increase in group-level performance in the inference by exclusion task was driven by a few individuals (mostly from the Chimpanzee B and the Orangutan groups) and not general learning effects.

As the second step, we investigated the relations between latent traits. That is, we

asked whether individuals with high abilities in one domain also have higher abilities in another. We fit pairwise LST models that modeled the correlation between latent traits between two tasks. In Phase 1, the only correlation that was reliably different from zero was that between quantity discrimination and inference by exclusion. In Phase 2, this finding was replicated and, in addition, four more correlations turned out to be substantial (see Figure 4B). One reason for this increase was the inclusion of the delay of gratification task. Across phases, correlations involving the gaze following task were the closest to zero, with quantity discrimination in Phase 2 being an exception. Taken together, the overall pattern of results suggests substantial shared variance between tasks – except for gaze following.

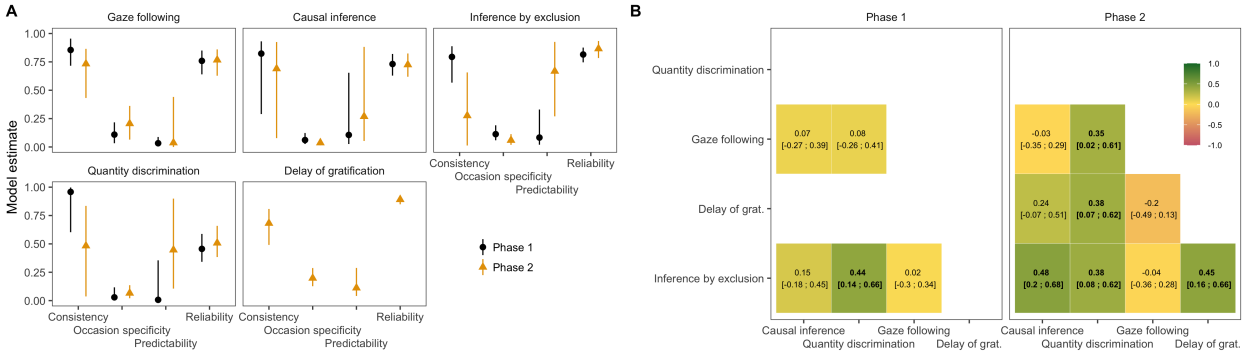

Figure 4. A. Estimates from latent state-trait model with autoregressive effecst for Phase 1 and 2. Consistency: proportion of (measurement-error free) variance in performance explained by stable trait differences. Occasion specificity: variance explained by variable states. Reliability: proportion of variance in raw scores explained by the trait and the state. B. Correlations between latent traits based on pairwise LST models between tasks with 95% Credible Interval. Bold correlations are reliably different from zero. The models for quantity discrimination and causal inference showed a poor fit and are not reported here (see supplementary material for details).

Predictability of individual differences

The results thus far suggest that individual differences originate from stable differences in cognitive abilities that might be shared between tasks. In the last set of analysis, we sought to explain the origins of these differences. That is, we analysed whether inter- or intra-individual variation in performance in the tasks could be predicted by variables that capture a) stable differences between individuals (group, age, sex, rearing history, time spent in research), b) differences that vary within and between individuals (rank, sickness, sociality), c) differences that vary with group membership (time spent outdoors, disturbances, life events), and d) differences in testing arrangements (presence of observers, study participation on the same day and since the last time point). We collected these predictor variables using a combination of directed observations and keeper questionnaires. This large set of potentially relevant predictors poses a variable selection problem. That is, we sought to find the minimal set of predictors that allowed us to accurately predict performance in the cognitive tasks. We chose the projection predictive inference approach because it provides and excellent trade-off between model complexity and accuracy (Pavone, Piironen, Bürkner, & Vehtari, 2020; Piironen, Paasiniemi, & Vehtari, 2020; Piironen & Vehtari, 2017). The outcome of this analysis is a ranking of the different predictors in terms of how important they are to predict performance in a given task. Furthermore, for each predictor, we get a qualitative assessment of whether it makes a substantial contribution to predicting performance in the task or not.

... but models also included a random intercept term, capturing variance specific to individuals not explained by the predictors.

Discussion

Summary

Individual differences are mostly stable across time. This matches well with the PPI

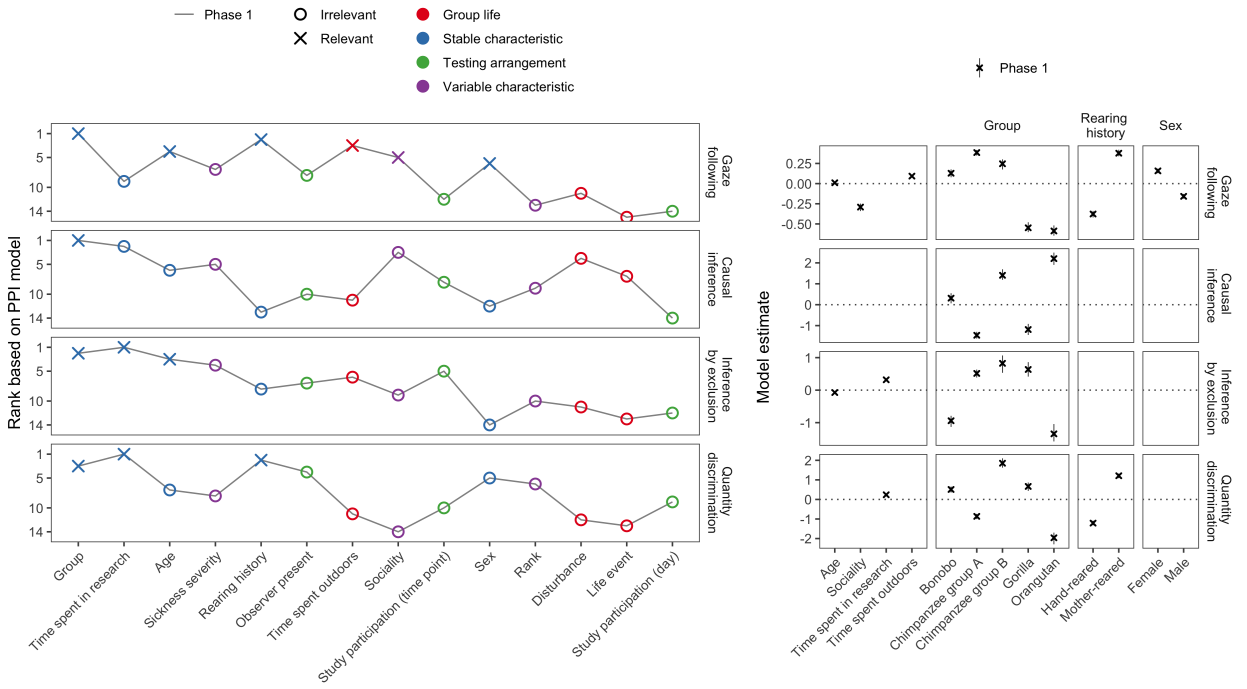

Figure 5. A. Latent state-trait model estimates for Phase 1 and 2. Consistency: proportion of (measurement-error free) variance in performance explained by stable trait differences. Occasion specificity: variance explained by variable states. Reliability: proportion of variance in raw scores explained by the trait and the state. B. Correlations between latent traits based on pairwise LST models between tasks with 95% Credible Interval. Bold correlations are reliably different from zero. The models for quantity discrimination and causal inference showed a poor fit and are not reported here (see supplementary material for details).

results selecting stable individual differences as the most important predictors. Group not species because A and B chimps not always align and few individual per group. But also the case that a large chunk of the variance remained unexplained. Most likely this reflects differential developmental effects.

Learned something about development. we saw change over time - but this change was slow, steady and not uniform. the hypothetical argument that looms large in studies of great ape cognition fearing associative learning effects that might account for findings is simply not supported by the data. The study presented ideal conditions for associative learning to happen (same tasks, with exact same order dozens of times), yet the change we saw was slow, steady - and not the same for everyone. No doubt apes can do pure associative learning – it is just very very slow.

Systematic relations between tasks. Where do they come from? same methods (cite christoph) or same cognitive mechanisms? larger samples needed to do this systematically and not just for 2 tasks at a time. collaboration (cite MP). Also Task based cognitive modelling would be good (cite RSApes, pragBat?)?

Methods

A detailed description of the methods and results can be found in the supplementary material available online.

Participants

Material

Procedure

Data analysis

References

Beran, M. J., & Hopkins, W. D. (2018). Self-control in chimpanzees relates to general intelligence. *Current Biology*, *28*(4), 574–579.

Cauchoix, M., Chow, P., Van Horik, J., Atance, C., Barbeau, E., Barragan-Jason, G., ... others. (2018). The repeatability of cognitive performance: A meta-analysis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1756), 20170281.

Coolidge, F. L., & Wynn, T. (2016). An introduction to cognitive archaeology. *Current Directions in Psychological Science*, *25*(6), 386–392.

Coqueugniot, H., Hublin, J.-J., Veillon, F., Houët, F., & Jacob, T. (2004). Early brain growth in homo erectus and implications for cognitive ability. *Nature*, *431*(7006), 299–302.

Currie, A., & Killin, A. (2019). From things to thinking: Cognitive archaeology. *Mind & Language*, *34*(2), 263–279.

Eid, M., Holtmann, J., Santangelo, P., & Ebner-Priemer, U. (2017). On the definition of latent-state-trait models with autoregressive effects: Insights from LST-r theory. *European Journal of Psychological Assessment*, *33*(4), 285.

Farrar, B., & Ostojic, L. (2019). *The illusion of science in comparative cognition.*

Gunz, P., Neubauer, S., Falk, D., Tafforeau, P., Le Cabec, A., Smith, T. M., ... Alemseged, Z. (2020). Australopithecus afarensis endocasts suggest ape-like brain organization and prolonged brain growth. *Science Advances*, *6*(14), eaaz4729.

Haslam, M., Hernandez-Aguilar, R. A., Proffitt, T., Arroyo, A., Falótico, T., Fragaszy, D., ... others. (2017). Primate archaeology evolves. *Nature Ecology & Evolution*, *1*(10), 1431–1437.

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186.

Herrmann, E., Call, J., Hernández-Lloreda, M. V., Hare, B., & Tomasello, M. (2007). Humans have evolved specialized skills of social cognition: The cultural intelligence hypothesis. *Science*, *317*(5843), 1360–1366.

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

Heyes, C. (2018). *Cognitive gadgets*. Harvard University Press.

Hopkins, W. D., Russell, J. L., & Schaeffer, J. (2014). Chimpanzee intelligence is heritable. *Current Biology*, *24*(14), 1649–1652.

Laland, K., & Seed, A. (2021). Understanding human cognitive uniqueness. *Annual Review of Psychology*, *72*, 689–716.

MacLean, E. L., Hare, B., Nunn, C. L., Addessi, E., Amici, F., Anderson, R. C., ... others. (2014). The evolution of self-control. *Proceedings of the National Academy of Sciences*, *111*(20), E2140–E2148.

MacLean, E. L., Matthews, L. J., Hare, B. A., Nunn, C. L., Anderson, R. C., Aureli, F., ... others. (2012). How does cognition evolve? Phylogenetic comparative psychology. *Animal Cognition*, *15*(2), 223–238.

ManyPrimates, Altschul, D. M., Beran, M. J., Bohn, M., Caspar, K. R., Fichtel, C., ... others. (2019). Collaborative open science as a way to reproducibility and new insights in primate cognition research. *Japanese Psychological Review*, *62*(103), 205–220.

Martins, E. P., & Martins, E. P. (1996). *Phylogenies and the comparative method in animal behavior*. Oxford University Press.

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

Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain*

Sciences, *31*(2), 109–130.

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

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: Evaluating the past, present and future of comparative psychometrics. *Animal Cognition*, *20*(6), 1003–1018.

Stevens, J. R. (2017). Replicability and reproducibility in comparative psychology. *Frontiers in Psychology*, *8*, 862.

Thornton, A., & Lukas, D. (2012). Individual variation in cognitive performance: Developmental and evolutionary perspectives. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1603), 2773–2783.

Tomasello, M. (2019). *Becoming human*. Harvard University Press.

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

Volter, C. J., Tinklenberg, B., Call, J., & Seed, A. M. (2018). Comparative psychometrics: Establishing what differs is central to understanding what evolves. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1756), 20170283.

Wobber, V., Herrmann, E., Hare, B., Wrangham, R., & Tomasello, M. (2014). Differences in the early cognitive development of children and great apes.

316 *Developmental Psychobiology*, *56*(3), 547–573.