A Longitudinal Study of Great Ape Cognition: Stability, Reliability and the Influence of Individual Characteristics

Anonymous CogSci submission

Abstract

Primate cognition research allows us to reconstruct the evolution of human cognition. However, temporal and contextual factors that induce variation in cognitive studies with great apes are poorly understood. Here we report on a longitudinal study where we repeatedly tested a comparatively large sample of great apes (N = 40) with the same set of cognitive measures. We investigated the stability of group-level results, the reliability of individual differences and the relation between cognitive performance and individual-level characteristics. We found results to be relatively stable on a group level. Some, but not all, tasks showed acceptable levels of reliability. Cognitive performance across tasks was not systematically related to any particular individual-level predictor. This study highlights the importance of methodological considerations — especially when studying individual differences – on the route to building a more robust science of primate cognitive evolution.

Keywords: Primate Cognition; Stability; Reliability; Individual Differences.

Introduction

Primate cognition research can inform us about the evolution of human cognition. This research has contributed significantly to our understanding of the shared and unique aspects of human cognition (Laland & Seed, 2021). But, like all other branches of cognitive science, primate cognition research faces some critical challenges: Because cognitive processes cannot be observed directly, they must be inferred from behavior. This kind of inference requires strong methods which specify the link between behavior and cognition. In this paper, we report on a longitudinal study that focuses on the stability, reliability and predictability of great apes' performance in a range of cognitive tasks.

To allow for generalization, study results need to replicate. That is, comparable results should be obtained when applying the same method to a new population of individuals. Psychological science has been riddled with problems of non-replicable results (Collaboration, 2015). Animal cognition research shows many of the characteristics that have been identified to yield a low replication rate in other psychological fields (Farrar, Voudouris, & Clayton, 2020; Stevens, 2017). Furthermore, replication attempts are rare in animal cognition research (Farrar, Boeckle, & Clayton, 2020). A recent review of experimental primate cognition research between 2014 and 2019 found that only 2 % of studies included a replication (ManyPrimates, Altschul, Beran, Bohn, Caspar, et al., 2019). Replications are rare, in part because researchers only have access to one sample of study participants and therefore cannot test a new sample. Nevertheless, in such conditions, we can ask a more fundamental question: how *repeatable* are the results of a study. That is if we test the same animals multiple times, do we get similar results? Repeatability could be seen as a pre-condition for replicability. In this study, we investigate the stability of results by repeatedly testing the same sample of great apes on the same tasks.

One way to explain cognitive evolution is to study how cognitive abilities cluster in different species. This approach needs reliable measures (Volter, Tinklenberg, Call, & Seed, 2018). Reliability refers to the stability of individual differences as opposed to group-level means. Reliability is paramount if a study's goal is to relate cognitive performance to individual characteristics or external variables: a measure cannot be stronger related to a second measure than to itself. Recent years have seen an increase of individual differences studies in animal cognition research (Shaw & Schmelz, 2017). In these studies, the reliability of the tasks is rarely assessed. Therefore, it is difficult to say if the absence of a relation between two variables is real or merely a consequence of low reliability. As part of this study, we investigate the re-test reliability of a range of commonly used cognitive tasks for great apes.

Researchers in animal cognition often assume that performance in cognitive tasks can (in part) be explained by individual-level characteristics such as age, sex, rank or rearing history. In many cases, such predictors are included without a specific hypothesis, either to control for potential effects or because they are implicitly assumed to influence cognitive performance in general. Habitually including these predictors without a theoretical indication is problematic because – in combination with selective reporting – it may increase the rate of false-positive results (Simmons, Nelson, & Simonsohn, 2011). As part of the study reported here, we investigated whether individual characteristics influence cognitive performance on a broad scale.

In the following, we describe the first results from a longitudinal study with great apes. We ask how stable performance is on a group level, how reliable individual differences are and to what extent these individual differences can be explained by a common set of predictors. We chose five tasks that cover a broad range of cognitive abilities: causal inference, inference by exclusion, gaze following, quantity dis-

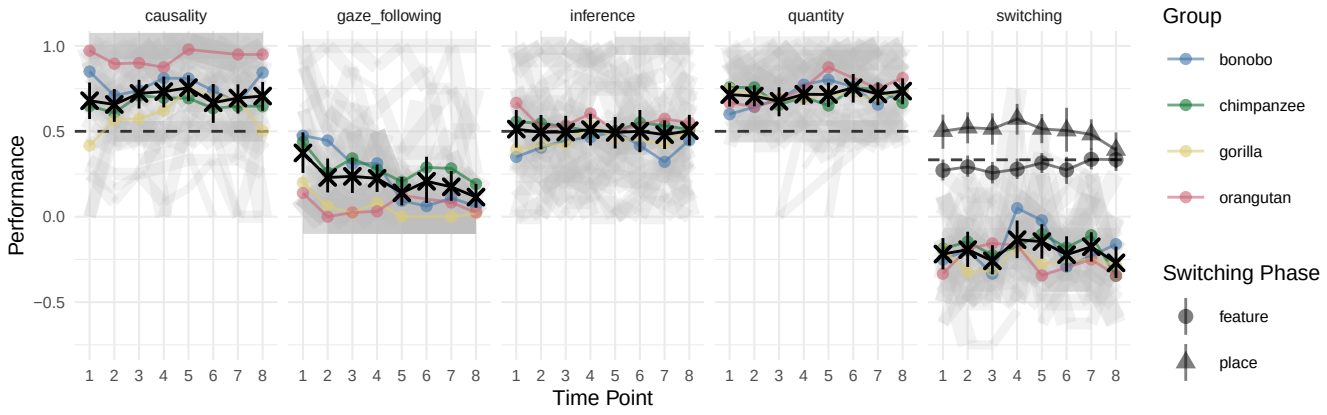

Figure 1: Results from the five cognitive tasks across time points. Black crosses show mean performance at each time point across species (with 95% CI). Colored dots show mean performance by species. Transparent grey lines connect individual performances across time points, with the line's width corresponding to the number of participants. Dashed line shows the chance level inference whenever applicable. The panel for switching includes triangles and dots showing the mean performance in the two phases from which the overall performance score was computed (see main text).

crimination, and switching flexibility. We tested a sample of individuals from four great ape species: Bonobos (*Pan paniscus*, Chimpanzees (*Pan troglodytes*), Gorillas (*Gorilla gorilla* and Orangutans(*Pongo abelii*) on regular intervals.

Methods

Participants

A total of 40 great apes participated at least once in one of the tasks. This included 8 Bonobos (3 females, age 7.3 to 38), 21 Chimpanzees (16 females, age 2.6 to 54.9), 6 Gorillas (4 females, age 2.7 to 21.6), and 6 Orangutans (4 females, age 17 to 40.2). The sample size at the different time points ranged from 22 to 38.

Apes were housed at the [masked for peer review]. They lived in groups, with one group per species and two chimpanzee groups. Research was noninvasive and strictly adhered to the legal requirements in Germany. Animal husbandry and research complied with the European Association of Zoos and Aquaria Minimum Standards for the Accommodation and Care of Animals in Zoos and Aquaria as well as the World Association of Zoos and Aquariums Ethical Guidelines for the Conduct of Research on Animals by Zoos and Aquariums. Participation was voluntary, all food was given in addition to the daily diet, and water was available ad libitum throughout the study. The study was approved by an internal ethics committee at the [masked for peer review].

Design, Setup and Procedure

We tested apes on the same five tasks every other week. Here we report the data from the first eight time points. The tasks were presented in the same order and with the same positioning and counterbalancing (to keep conditions constant between individuals and across occasions). Apes were tested in familiar sleeping or observation rooms by a single experimenter. Whenever possible, they were tested individually. For each individual, the tasks at one time point were usually spread out across two consecutive days with causality and inference on day 1 and quantity and switching on day 2. Gaze following trials were run at the beginning and the end of each day. The basic setup comprised a sliding table positioned in front of a clear Plexiglas panel with three holes in it. The experimenter sat on a small stool and used an occluder to cover the sliding table.

Causality The causality and inference tasks were modeled after (Call, 2004). Two identical cups with a lid were placed left and right on the table. The experimenter covered the table with the occluder, retrieved a piece of food, showed it to the ape, and hid it in one the cups outside the participant's view. Next, they removed the occluder, picked up the baited cup and shook it three times, which produced a rattling sound. Next, the cup was put back in place, the sliding table pushed forwards, and the participant made a choice by pointing to one of the cups. If they picked the baited cup, their choice was coded as correct, and they received the reward. If they chose the other cup, they did not. On each time point, participants received 12 trials.

Inference Inference trials were identical to causality trials, but instead of shaking the baited cup, the experimenter shook the empty cup. On each time point, participants received 12 trials. Inference trials were intermixed with causality trials.

Gaze Following The gaze following task was modeled after (Brauer, Call, & Tomasello, 2005). The experimenter sat opposite the ape and handed over food at a constant pace. That is, the experimenter picked up a piece of food, briefly held it out in front of her face and then handed it over to the participant. At some point, the experimenter looked up (i.e., moving her head up) while holding up the food in front of her head. After 10s, the experimenter looked down again and handed over the food. We coded whether the subject looked

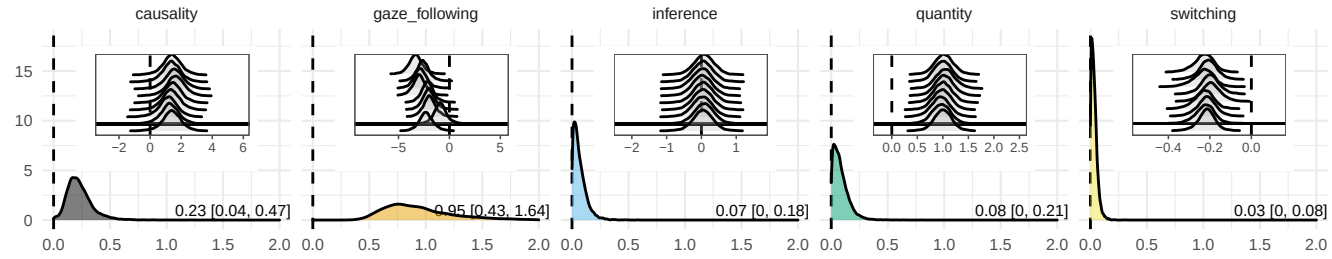

Figure 2: Posterior distributions for τ from the meta-analytic models for each task. Numbers denote mean and 95% HDI for τ. Insets show the posterior distribution for the model intercept estimate at each time point and the overall estimate at the bottom (separated by the black line).

up during the 10s interval. Participants received a total of 8 trials, spread out across the two test days.

Quantity Discrimination For this task, we followed the general procedure of (Hanus & Call, 2007). Two small plates were presented left and right on the table. The experimenter placed 5 small food pieces on one plate and 7 on the other. Then they pushed the sliding table forwards, and the subject made a choice. We coded as correct when the subject chose the plate with the larger quantity. There were 12 trials per time point.

Switching This task was modeled after (Haun, Call, Janzen, & Levinson, 2006). Three differently looking cups (metal cup with handle, red plastic ice cone, red cup without handle) were placed next to each other on the table. There were two conditions. In the place condition, the experimenter hid a piece of food under one of the cups in full view of the participant. Next, the cups were covered by the occluder and the experimenter switched the position of two cups, while the reward remained in the same location. We coded as correct if the participant chose the location where the food was hidden. Participants received four trials in this condition. The place condition was run first. The feature condition followed the same procedure, but now the experimenter also moved the reward when switching the cups. The switch between conditions happened without informing the participant in any way. A correct choice in this condition meant choosing the location to which the cup plus the food were moved. Here, participants received eight trials. The dependent measure of interest for this task was calculated as: `[proportion correct place] - (1 - [proportion correct feature])`. Positive values in this score mean that participants could quickly switch from choosing based on location to choosing based on feature. High negative values suggest that participants did not or hardly switch strategies.

Analysis and Results

We combined the data from all species for the analysis because sample sizes for some species were too small to get representative estimates. However, we accounted for the nesting of subjects in species as part of the random effect structure of our models. All analyses were run in `R` (R Core Team, 2018). Bayesian multilevel models were implemented using the package `brms` (Burkner, 2017) and default priors. All models included random intercepts for participants nested within group and random slopes for trial (`trial|group/subject`). Data and analysis code can be found in the associate online repository (see below).

Stability

First, we looked at group-level stability in performance. That is, we asked how much performance varied across time points in the different tasks. For this analysis, we ignored the temporal order of the different time points and treated them as repetitions of the same experiment (i.e., time point was treated as a factor instead of a numerical variable). As such, we asked a meta-analytic question: how much variation is there between different instances of the same experiment? To answer this, we fitted a mixed model with a random intercept term for time point to the data from each task[1]. As part of each model, we estimated a standard deviation of the random intercept term (τ), which reflects the variation between time points.

Figure 1 visualizes performance across time points. For causality, inference and quantity, we can evaluate group-level performance by comparing it to chance (50% correct = intercept of 0 in link space). Group-level performance was reliably above chance for causality and quantity but at chance for inference. There is no such reference level for gaze following, and we can simply say that at least some individuals of all species followed the experimenter's gaze. The switching score was consistently negative, suggesting that - on a group level - apes did not switch strategies.

Figure 2 shows the posterior distribution of τ for each task. While performance was very stable for inference, quantity and switching (τ very close to 0), performance was slightly more variable for causality and varied substantially for gaze following. For causality, variation did not seem to follow a clear temporal pattern. On the other hand, for gaze following, there seems to be a downward trend with apes (as a group) becoming less likely to follow the experimenter's gaze. We

[1] We modeled the trial by trial data using a binomial distribution in a logistic GLMM for all tasks, except switching. Here we modeled the score (by time point) as a truncated normal distribution. As mentioned above, these models included random intercept terms for individuals nested within groups

explore this temporal pattern in more detail below. Taken together, we may say that 4 out of 5 measures yield stable measures of group-level performance. For inference, however, stability corresponds to a stable performance at chance level, which suggests that the task was rather difficult. Whether that meant that participants simply guessed on each trial, we will explore in the next section.

Reliability

Next, we asked how stable performance was on an individual level. This question also relates to each task's reliability - how well suited it is to capture differences between individuals. In general, reliability is high if individuals are consistently ranked across measurement instances. One way to assess reliability is to correlate performance from two time points (re-test reliability). Because we had multiple time points, we computed pairwise correlations for all combinations of time points (total of 28 unique correlations per task). This resulted in a distribution of correlations, which we visualize in Figure 3. Results suggest good re-test reliability for gaze following, causality and inference, variable reliability for quantity and poor reliability for switching. This pattern is interesting in light of the group-level performance we reported above: stable performance on a group level (stability) does not imply stable individual differences (reliability). We come back to this point in the discussion.

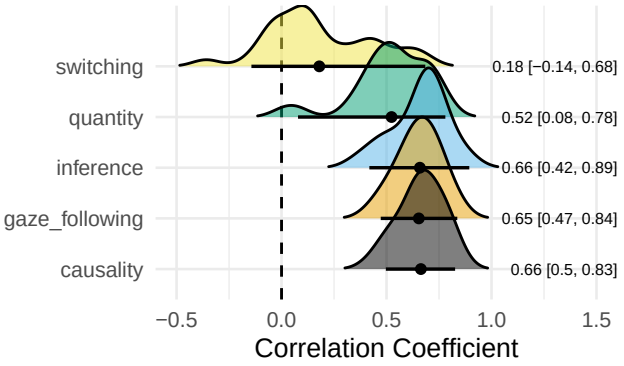

Figure 3: Distribution of correlations between time points for each task. Dots represent the mean of the distribution with 95% HDI. Numbers denote mean and 95% HDI.

Predictors

In the final set of analysis, we investigated if variation in cognitive performance could – in part – be explained by participant characteristics. We chose to look at variables that are commonly analyzed in the primate cognition literature: age, sex, rank and rearing history. Rank was rated by animal keepers at every time point, and rearing history was classified as "mother reared," "human reared" or "unknown."

For each task, we ran the same five models[2]: A baseline model predicting performance by time point (numerical) and trial as well as four models, each with one of the predictors (age, sex, rank and rearing history), added to the baseline model. We did not investigate any interaction models (interactions among the predictors or with time point) because we had no specific hypothesis in that direction. We used Bayesian model comparison based on WAIC (widely applicable information criterion) scores and weights (McElreath, 2016). This comparison tells us which of the models considered makes the best out-of-sample predictions. If the model with one predictor (e.g., age) were consistently assigned the highest weight across tasks, we would conclude that participants' age best predicts cognitive performance.

Table 1 gives WAIC scores and weights for each model and task. Figure 4 shows the posterior distribution of the test predictors (as well as for time point). The baseline model was ranked highest across tasks (first or second for all tasks), suggesting that none of the test predictors was consistently related to performance. Within the baseline model, the estimate for time point was close to 0 for all tasks except gaze following, for which it was mostly negative (reflecting the downward trend we saw in Figure 1).

For gaze following, the model including sex as a predictor was ranked highest: males were somewhat less likely to follow the experimenter's gaze. For quantity and switching, the rank model was rated highest with lower-ranking individuals showing better quantity discrimination or switching abilities. In the case of switching, however, the model results should be interpreted with caution. The low re-test correlations suggest that the task does not reliably measure the cognitive ability in question. Thus, the variation in performance that the model tries to explain might not have a cognitive origin and could equally well be due to factors we did not capture.

Discussion

We tested the same sample of great apes repeatedly on five cognitive tasks. This design allowed us to address some pressing questions in primate cognition research: How stable is group-level performance in cognitive tasks? How reliable are the results of these tasks? How much do individual characteristics influence performance? Below we discuss the results in light of these questions.

Performance was relatively stable for all tasks except gaze following. This result is somewhat surprising given that individuals were differentially reinforced in all tasks – except gaze following. Furthermore, counterbalancing and positioning were exactly the same at each time point. Together, this creates a potentially ideal learning scenario. How can we interpret this lack of improvement in the tasks other than gaze following? One explanation could be that the different routes to solving the task constitute incompatible information sources. For example, in the case of causality, apes could spontaneously solve the task by inferring that the food caused the sound. Alternatively, they could learn that food is under the cup the experimenter touches whenever they hear a rattling sound. In principle, these two information sources

[2]We used the same response distributions as in the stability analysis.

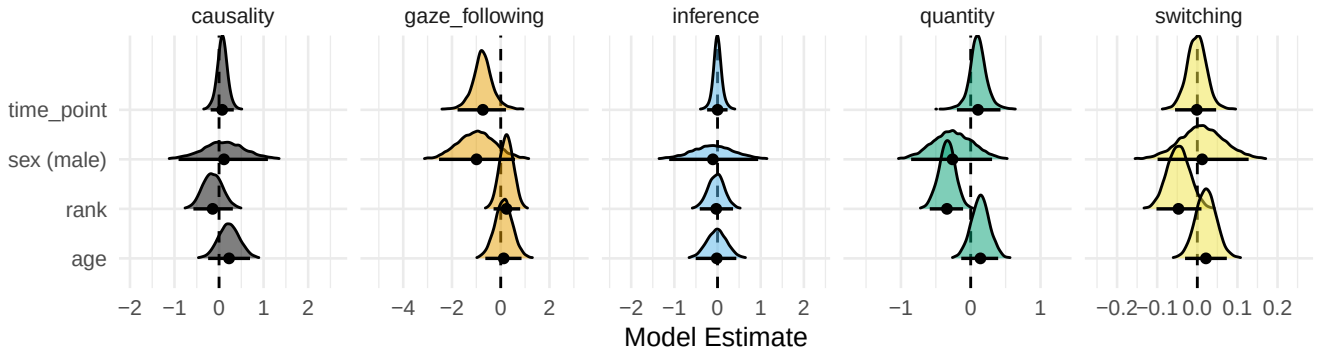

Figure 4: Posterior distribution for the test predictors for each task. Dots represent the mean of the distribution with 95% HDI. Samples for time point are drawn from the baseline model. For all models except switching, the estimates are given in link space. No samples are shown for the rearing model.

could easily be integrated and supplement one another, resulting in improved performance over time. The absence of improvement could mean that apes rely on spontaneous inferences alone, thereby ignoring repeating contingencies. However, many alternative explanations are possible. For example, many apes in [masked] have had years of experience with the kind of tasks we included in the study. Thus, the absence of improvement might indicate that they already reached an individual performance maximum. The continuation of this project might help to shed light on these questions. For now, we may conclude that short term improvements based on learned arbitrary relations are unlikely to occur in great apes. In support of this, when primates learned arbitrary relations in previous studies, it typically took a very long time and an elaborate training regime (e.g. Allritz, Call, & Borkenau, 2016).

Three out of five tasks showed acceptable levels of reliability. Importantly, reliability is independent of group-level performance (leaving aside floor and ceiling effects) (see Hedge, Powell, & Sumner, 2018). Here, we see such a pattern for inference: Group-level performance was consistently at chance level for every time point. On a group level, one would conclude that great apes did not make the inference in question. However, the task was highly reliable, suggesting that it accurately captured individual differences. Together with the observation that some individuals consistently performed at ceiling (see grey transparent lines in Figure 1), this suggests that the task is well suited to measure inferential abilities on an *individual* level. The opposite pattern holds for quantity. Here, group-level performance was consistently above chance, but individual differences were not very consistent. This suggests that variation was due to sources other than systematic differences between individuals. This phenomenon is quite common in the human adult cognitive literature (Hedge, Powell, & Sumner, 2018). It arises when experimental tasks (optimized for low variance in measurement) are used to study individual differences (requiring high variance in measurement). Taken together, we may recommend that researchers investigate the psychometric properties of an experimental task before they use it to study individual differences. When planning to study individual differences by relating measures to one another, researchers might be well advised to first study the reliability of these measures. Even though this takes considerable time and effort, it increases the chances of finding meaningful effects.

We did not find that one of the individual-level characteristics (age, sex, rank or rearing history) was consistently related to performance across tasks. A baseline model, predicting performance by time variables alone, was, on average, rated highest in the different model comparisons. The model including rank was rated highest for two tasks (quantity and switching). However, in the case of switching, this should be interpreted with caution in the light of low reliability of the task (see results section). Moving forward, we will explore additional predictors, to see if we do find some that are related to cognitive performance more broadly. For now, we may conclude that researchers should carefully select predictors based on theoretical considerations. Including them as a default or to control for potential effects might make models unnecessarily complex – and might not even have the desired effect (see Westfall & Yarkoni, 2016).

For our analysis, we combined the data from all species, neglecting potential species differences. The reason is that the sample size for each species was too small to really differentiate individual- from species-level differences. This is a common problem in primate cognition research. Species-level inferences require data sets that are beyond the resources of individual labs. A promising way forward to overcome this limitation is the *ManyPrimates* project; a large-scale collaborative initiative established to create an infrastructure to support the pooling of resources across labs (ManyPrimates, Altschul, Beran, Bohn, Call, et al., 2019).

The data we have reported here are the first couple of waves in a longitudinal study which we hope to continue for at least one year. As part of it, we will record additional variables that might explain variation in cognitive performance such as social network data or live history variables (sickness, birth and death of group members, etc.). We hope that this project

will contribute to our understanding of the dynamic nature of primate cognition.

Task	Model	WAIC	SE	Weight
Causality	baseline	2432.35	52.98	0.25
	rank	2432.80	53.03	0.20
	age	2433.05	53.09	0.18
	sex	2432.56	53.07	0.23
	rearing	2433.45	53.09	0.15
Gaze following	baseline	1133.68	50.19	0.22
	rank	1134.56	50.26	0.14
	age	1134.31	50.29	0.16
	sex	1132.74	50.19	0.35
	rearing	1134.65	50.41	0.13
Inference	baseline	2915.33	44.01	0.24
	rank	2916.23	44.03	0.16
	age	2915.98	44.07	0.18
	sex	2915.19	44.13	0.26
	rearing	2916.17	44.20	0.16
Quantity	baseline	2501.47	47.63	0.23
	rank	2500.65	47.74	0.35
	age	2502.04	47.73	0.18
	sex	2502.54	47.70	0.14
	rearing	2503.18	47.75	0.10
Switching	baseline	25.56	22.13	0.30
	rank	25.12	21.94	0.37
	age	27.20	22.27	0.13
	sex	27.19	22.14	0.13
	rearing	28.41	22.31	0.07

Table 1: WAIC Scores and weights for each predictor model and task.

Corresponding data and code are available at [masked for peer review]

Acknowledgements

Masked for peer review

References

Allritz, M., Call, J., & Borkenau, P. (2016). How chimpanzees (pan troglodytes) perform in a modified emotional stroop task. *Animal Cognition*, *19*(3), 435–449.

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

Burkner, P.-C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28.

Call, J. (2004). Inferences about the location of food in the great apes (pan paniscus, pan troglodytes, gorilla gorilla, and pongo pygmaeus). *Journal of Comparative Psychology*, *118*(2), 232.

Collaboration, O. S. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251).

Farrar, B., Boeckle, M., & Clayton, N. (2020). Replications in comparative cognition: What should we expect and how can we improve? *Animal Behavior and Cognition*, *7*(1), 1.

Farrar, B., Voudouris, K., & Clayton, N. (2020). Replications, comparisons, sampling and the problem of representativeness in animal behavior and cognition research.

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

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

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.

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

ManyPrimates, Altschul, D. M., Beran, M. J., Bohn, M., Call, J., DeTroy, S., ... others. (2019). Establishing an infrastructure for collaboration in primate cognition research. *PLoS One*, *14*(10), e0223675.

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.

McElreath, R. (2016). *Statistical rethinking: A bayesian course with examples in R and Stan* (pp. xvii, 469). Boca Raton: CRC Press.

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

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.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366.

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

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.

Westfall, J., & Yarkoni, T. (2016). Statistically controlling for confounding constructs is harder than you think. *PloS One*, *11*(3), e0152719.