Replication and reproducibility in primate cognition research.

Farrar, Benjamin; Krupenye, Christopher; Motes-Rodrigo, Alba; Tennie, Claudio; Fisher, Julia; Altschul, Drew; Ostojić, Ljerka

Replication and Reproducibility in Primate Cognition Research

Benjamin G. Farrar[1, 2], Christopher Krupenye[3], Alba Motes-Rodrigo[4], Claudio Tennie[4], Julia Fischer[5], Drew M. Altschul[6], Ljerka Ostojić[7]

[1]Department of Psychology, University of Cambridge, Downing Street, Cambridge, CB2 3EB, UK

[2]Institute for Globally Distributed Open Research and Education (IGDORE)

[3]Department of Psychological & Brain Sciences, Johns Hopkins University Baltimore, MD, 21218, USA

[4]Department of Early Prehistory and Quaternary Ecology, University of Tübingen, Schloss Hohentübingen, Burgsteige 11, 72070 Tübingen, Germany

[5]Cognitive Ethology Laboratory, German Primate Center/University of Göttingen, Göttingen, Germany

[6]Department of Psychology School of Philosophy, Psychology and Language Sciences, The University of Edinburgh Edinburgh, UK

[7]Department of Psychology, Faculty of Humanities and Social Sciences, University of Rijeka, Sveučilišna avenija 4, Rijeka, Croatia

Abstract

Replication is an important tool used to test and develop scientific theories. Areas of biomedical and psychological research have experienced a replication crisis, in which many published findings failed to replicate. Following this, many other scientific disciplines have been interested in the robustness of their own findings. This chapter examines replication in primate cognitive studies. First, it discusses the frequency and success of replication studies in primate cognition and explores the challenges researchers face when designing and interpreting replication studies across the wide range of research designs used across the field. Next, it discusses the type of research that can probe the robustness of published findings, especially when replication studies are difficult to perform. The chapter concludes with a discussion of different roles that replication can have in primate cognition research.

Keywords: Generalizability, Heterogeneity, Primates, Replication, Reproducibility, Research Design, Sampling

Replication and Reproducibility in Primate Cognition Research

Most primate species are threatened or endangered in their natural habitats, and many are as rare in captivity as they are in the wild (Estrada et al., 2017). Of known populations, only a portion are available for research: many wild populations are unhabituated, many captive facilities do not prioritize research, and certain research protocols – like touchscreen paradigms – require extensive training before animals can participate. For scientists, this results in a landscape of relatively small study populations, where independent replications are infrequent and centered around a few species (e.g., chimpanzees, rhesus macaques, capuchin monkeys). This landscape contrasts with textbook views of "the scientific method," in which many groups of scientists work independently in very similar environments, replicating their own and others' work in order to ensure the accuracy of scientific progress (Dunlap, 1926; Zwaan et al., 2018). In contrast to this supposed scientific ideal, many experiments in primate cognition will never be replicated, posing a problem for cumulative science in the field. In this chapter, we outline the current status of replication in primate cognitive studies and discuss the challenges of replication in a field with limited resources and unique samples.

Goals of replication

Many terms are used when discussing the replicability of scientific results (such as reliability, repeatability or reproducibility), without consensus on when each term should be used. For the purpose of this chapter, we will use two terms: replication and robustness. *Replication* refers to the process of carrying out an experiment that matches some elements of a previous experiment. If such a study produces results that are statistically consistent with a previous experiment, the effect in question is likely *robust* to sampling variance and thus has replicated "successfully."

Researchers usually conduct replication studies with the goal of assessing the credibility of a scientific finding or its suitability as the basis of future research (Fidler & Wilcox, 2018; Fraser et al., 2020). In some studies, researchers closely match all elements of an original study, conducting what is known as a direct replication study. Alternatively, researchers may closely replicate some elements of a study, but vary others, in order to test a claim in a novel way or to assess its generalizability. Such studies are often labelled as conceptual replications, replications and extensions (if they involve novel stimuli, Beran, 2018), "quasi-replications" or between-site and between-species comparisons if the study is performed in a novel group, species or site (Farrar et al., 2020; Palmer, 2000). Terms such as direct and conceptual replication are useful in highlighting the researcher's intentions when performing a replication study. However, they should not be treated dichotomously; each replication study falls on a spectrum between direct and conceptual. At the most direct end of the spectrum,

researchers might repeat the same protocol used by a previous study on the same animals that were previously tested. At the most conceptual end, researchers might test a new sample of a different species in a new setting using a task that has been modified from a previous study.

A replication crisis

Interest in scientific replication surged in the 2010s. Following the claim that most published research findings are false (Ioannidis, 2005), large-scale replication projects in biomedical and psychological disciplines found that many studies could not be replicated successfully. In psychology, only 36% of 97 replication studies returned significant results in the same direction as the original study (Open Science Collaboration, 2015), and only 14 of 21 social science studies selected from *Nature* and *Science* replicated successfully (Camerer et al., 2018). In biomedical research, less than half of a sample of published preclinical cancer research findings replicated successfully (Begley & Ellis, 2012; Prinz et al., 2011), and in ecology, the materials necessary to simply reproduce a previous study's analysis (useable data and code) are available in less than a quarter of papers (Culina et al., 2020; Minocher et al., 2020). Such a low results replicability and computational reproducibility of research findings have contributed to some areas of science being in "crisis". The extent to which primate cognition research faces a similar crisis is currently unknown, and replication studies will be key in assessing this. Fortunately, the field of replication research has developed theoretically and practically over the past 10-years. By understanding the many causes and definitions of low replicability (Fiedler & Prager, 2018; Schauer & Hedges, 2020), and how and when to perform strong replication studies (Alexander & Moors, 2018; Camerer et al., 2018; Field et al., 2019; Machery, 2020; Schauer & Hedges, 2020; Smith & Little, 2018), primate cognition research may be able to avoid the shock that may follows when discovering that most findings in a given field cannot be easily replicated.

The status of replication in primate cognition research

Primate cognition researchers are exposed to the same incentives that have produced irreplicable findings in other fields — a pressure to produce many novel and appealing findings from limited data. Consequently, primate cognition researchers should expect many findings to be difficult to replicate, especially results coming from low power research with a strong publication bias (Farrar et al., 2020). However, primate cognition research is heterogeneous, and the field will contain many robust findings, too. These findings will often stem from studies that use hundreds or thousands of trials

within individual animals (Skinner, 1956; Smith & Little, 2018; Zwaan et al., 2018). However, little formal research has been conducted into the frequency, success and likelihood of replication in primate cognition, which we now survey.

How frequent are replications performed in primate cognition?

According to data compiled by the ManyPrimates project (2019), only 8.7% (50/574) of primate cognition studies published from January 2014 to October 2019 were replication studies, defined as studies that tested different populations of the same species with the same methodology (i.e., direct replications). Notably, less than one percent (0.6%, 4/574) of studies were within-paper replications, in which the authors conducted and reported replication studies within an individual publication (for examples, see Forss et al., 2020; Krupenye et al., 2016; Wallace et al., 2017). These data suggest that direct replication is not a routine aspect of primate cognition research. However, the rate of replication likely differs between study designs, laboratories and individual researchers. Due to this heterogeneity, it is difficult to interpret what the 8.7% rate of direct replication studies means. Does it mean that every experiment has an 8.7% chance of being replicated? Probably not. Most likely the 8.7% figure illustrates that replication studies in primate cognition likely come from a few simple tasks being replicated many times, often by a minority of laboratories.

In contrast to direct replication studies, conceptual replications appear to be common in primate cognition, although there are no published estimates of their exact frequency. In theory of mind research, for example, researchers have employed many different, but conceptually similar, study designs in order to test the same claims (Halina, 2020). However, a high rate of conceptual replication does not necessarily protect research fields from a replication crisis. Conceptual replications rarely put the original findings at risk because inconsistent results can be explained away by differences between the two studies. Moreover, if successful conceptual replications are more likely to be published than unsuccessful ones, a high rate of conceptual replication can actually lead to false findings being reinforced in the literature (Nissen et al., 2016).

Testing replication success in primate cognition

Understanding the robustness of core findings in primate cognition requires some amount of direct replication. However, performing research in primate cognition is costly, and large-scale replication projects like those performed in human psychological research are infeasible. As such, researchers may wish to also focus on indirect assessments of replicability. For individual studies, the reproducibility of a statistical analysis can be assessed by re-performing it (Culina et al., 2020). Furthermore, researchers can assess the strength of evidence of a statistical effect through investigating statistical markers, such as *p*-values and uncertainty intervals (Francis, 2014). More general information about the robustness of a body of research can come through programs of meta-

research. This research could entail estimating the prevalence of publication bias across a set of studies (Scheel et al., 2020) or using meta-analytic methods to quantify the overall strength of evidence for effects (e.g. Simonsohn et al., 2014). In addition to quantitative analyses, researchers may opt for qualitative measures. For example, surveys, interviews and ethnographies of various stakeholders (researchers, reviewers, funders and editors) can help to build up a stronger picture of the research and publication practices that produce the primate cognition literature (Candea, 2013; Fraser et al., 2018, 2020; Neuliep & Crandall, 1990; Peterson, 2016). However, just like any other scientific process, these tools must first be developed and validated, and it is unlikely that they will provide immediate answers regarding the replicability of primate cognition research. Quantitative approaches, in particular, will suffer from low statistical power and excessive heterogeneity themselves (Farrar et al., 2020), making a mixed-methods approach to understanding replicability in primate cognition crucial.

How do primate cognition researchers choose what to replicate?

When choosing whether, when and what to replicate, researchers must assess the costs and benefits of performing the replication study. This will be influenced by scientific interests (e.g., the potential of the replication study to increase understanding), personal interests (e.g., the amount of effort required and possibility of publication) and ethical regulations, as well as being constrained by resources and funding availability. Because of this, certain tasks in certain scenarios may be more likely to be replicated than others, and there may be a conflict between what should be replicated from a scientific perspective, and what is actually replicated.

Notably, tasks that are low cost, quick and adaptable may be more likely to be replicated across time and sites than those that are expensive, slow and difficult to adapt. Accordingly, tasks using few and simple apparatuses with little training requirements appear to be those that are replicated most often. For example, tasks using simple tube apparatuses, such as trap-tube tasks and tube tasks for handedness, make up a significant proportion of replication studies in primates. In the tube task, which has been replicated several times both within and between species (Chapelain et al., 2011; Hopkins et al., 2004; Llorente et al., 2011; Motes Rodrigo et al., 2018; Nelson et al., 2015), a tube with two openings and a food reward smeared in the middle is provided to the test subjects in order to assess which hand they use to retrieve the food. As such, the task can be applied easily and repeatedly to nearly all testable primate groups.

A positive feedback loop may then exist whereby tasks that are used in replication studies are more likely to be subject to further replication attempts. As more and more data become available on certain tasks, interpretative frameworks can be built around them and researchers can more easily produce a narrative around their data; more comparisons are possible, the results are more easily contextualized,

and less work may be needed to justify the task design (Latour & Woolgar, 1986). This feedback loop may partly explain the frequency at which test-batteries (Herrmann et al., 2007; Schmitt et al., 2012), tests assessing animals' reactions to mirrors (Anderson & Gallup, 2011), inhibition tests (MacLean et al., 2014), tests of spatial memory with arrays of cups (Many Primates et al., 2019), and quantity judgement tests using food sets (Beran, 2001) are replicated across not just primates, but animals more broadly (e.g., Brecht et al., 2020; Krasheninnikova et al., 2019).

While biases towards replicating simpler and more popular tasks may exist, this is not necessarily detrimental for scientific progress. In fact, focusing replication attempts on tasks that are easy to perform can be justified from both the perspective of productivity and informativeness (see Krasheninnikova et al., 2020, for the case of test batteries). Simpler tasks allow researchers to collect more data from more samples of animals, which then allow for more comparisons to be made (Beach, 1950). Simpler tasks can also allow data that are representative of heterogeneous populations to be generated more easily than complex tasks. However, prolifically replicating simple tasks can also produce weak research findings (Barrett, 2015; Eaton et al., 2018; Farrar et al., 2020; Scheel et al., 2016), particularly if they are made at the expense of mechanistic understanding. Rather than collecting data from many different animals, strong tests of scientific theories can be, and are, made using small samples of unique animals (Craig & Abramson, 2018; Leonelli, 2018; Smith & Little, 2018).

Challenges When Performing and Interpreting Replication Studies with Primates

Inaccessible Samples

The largest impediment to performing replications in primate cognition is having access to the primates of interest. Many researchers who would want to perform replication studies will be unable to do so simply because they do not have access to primates. However, even when researchers have access to primates, samples are often limited to their own or partnering facilities, like a local zoo. These settings are not standardized and vary across a range of variables that affect the behaviour of the test subjects they house. For example, primates of the same species will vary between sites in their genetics, age, diet, space, enrichment, and sociality. Such variables are often unmeasured, which makes it difficult to precisely match the methods of a previous study at a new site. As a result, direct replication studies can be hard to perform. In extreme cases, direct replication studies in independent samples are impossible (Leonelli, 2018), which is the case where unique samples are studied (e.g., with uniquely trained primates, or when investigating a population-specific behaviour or a case study of a rare event).

Between-Site Variation in Behaviour

Unique samples notwithstanding, many experiments in primate cognition can be replicated in some form. Yet, primates are notorious for their flexible behaviour. When working with primates, it is hard to engineer experimental contexts that consistently elicit, even from the same individual, similar behaviour or a single cognitive strategy. No matter how elegant the design, most studies impose cognitive, perceptual, motivational, and physical demands on subjects that are necessary to measure the construct of interest but may be independent of it (Colombo & Scarf, 2020; Mendes et al., 2011; Rowe & Healy, 2014; Schubiger et al., 2016, 2020). This can be exacerbated both within and across individuals by changes in mood, motivation and context that can heavily impact performance.

Because primate behaviour varies considerably across space and time, replication studies in primate cognition should not be expected to always return numerically similar results to an original study. This is especially the case when replication studies are performed at different sites, as the primates at each site will differ systematically from each other (i.e., primates will be more similar to other primates within their own site than at different sites, on average). This poses a problem for interpreting replication studies that yield contrasting results (Farrar et al., 2020a; Farrar et al., 2020b), particularly when the replication studies are between-species. To highlight this, consider a multi-laboratory multi-species study that recorded the performance on the cylinder and the A-not-B tasks in many primate species (MacLean et al., 2014). In the cylinder task, after habituation, animals are presented with a transparent tube containing a food reward. To obtain the reward, the animals must inhibit their tendency to directly approach the food, and instead detour to one of the tube's openings to gain access. Figure 1 displays the between-site performance of four primate species in this task (squirrel monkeys, orangutans, gorillas and capuchin monkeys). The groups of orangutans and capuchin monkeys performed close to ceiling at both sites, and a similar pattern may hold for the small number of gorillas. However, in contrast, the squirrel monkeys' behaviour differed markedly between sites: zoo-housed squirrel monkeys tested by University of St Andrews researchers had a median of 6 out of 10 correct choices, whereas lab-housed squirrel monkeys tested by Kyoto University researchers had a median of 0.5 out of 10 correct choices.

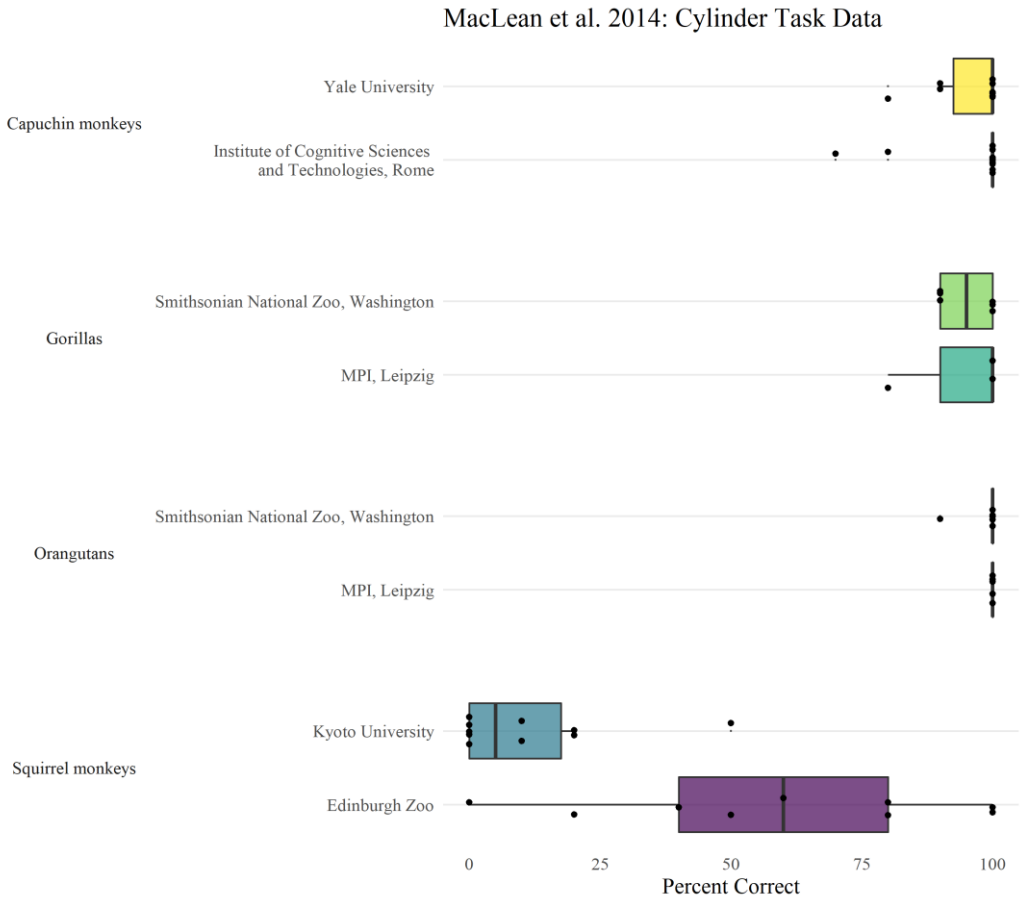

Figure 1: The performance of four primate species across two different sites on the cylinder task. Data from MacLean et al. 2014, and visualisation adapted from Farrar et al. 2020.

Clearly, we should not expect replication studies on one sample of primates to always produce the same numerical result as another sample of primates, even of the same species. However, while between-site variation may often modulate the magnitude of certain effects, it is unclear whether between-site variation is large enough to affect the *presence* or *direction* of psychological effects. For example, while diet may influence cognition across animals (Murphy et al., 2014), the ability of a primate species to learn to pass a given cognitive test should not be influenced by minor changes in diet. Provided studies are designed with sufficient statistical power, researchers should expect robust effects to replicate successfully across sites, where success is defined by the direction, rather than the absolute magnitude of an effect.

Defining Replication Success

Deciding whether a replication is successful or not can be a difficult process, and usually consists of two stages, i) asking whether the results of the replication study are statistically consistent with the results of the original study, and, ii) asking what the theoretical implications of such consistency are

(Nosek & Errington, 2020). Statistical significance can be used as a marker of statistical consistency (Open Science Collaboration, 2015), particularly when focusing on the direction of a single effect. However, statistical significance alone can be misleading: for example, a replication study yielding $p = 0.051$ is barely different from a replication study yielding $p = 0.049$. Similarly, if both original and replication studies have low power to detect theoretically important effect sizes, replication studies will return non-significant results most of the time, even if the underlying effect is robust (Schauer & Hedges, 2020a).

Nevertheless, some replication studies will provide strong evidence to support a claim, others will be ambiguous, and others will offer convincing counterevidence to an original claim. When an independent replication study affirms a previous claim, we learn that the effect in question is robust to changes in settings and subjects. If the replication is non-independent, i.e., being performed by the same experimenters, on the same animals, or being affected by similar bias, we learn less about the robustness of the effect (Ioannidis, 2012). In contrast, when replications do not affirm the previous claim, researcher must decide how much their confidence in the original claim changes. This requires a quality assessment of both the original study and the replication study, as well as an evaluation of each study's ability to test the claim at hand. If the original study had a stronger design than the replication study, the results of the replication study need not be weighted strongly. However, if the replication study improved on the design of the original, for example by increasing statistical power or implementing procedures to reduce bias (e.g., blinding or pre-specified analysis plans), greater weighting should be given to the replication study's results. Finally, it is possible to combine the results of multiple studies through meta-analysis. This is a strong procedure when the quality of each study can be guaranteed. However, any bias in the studies included in the meta-analysis will similarly bias the meta-analytic estimates.

Replication studies in the field

Laboratory studies are often contrasted with field studies in primate cognition. While greater standardization might be possible in laboratory studies, the challenges interpreting and performing replication studies in the field are similar, particularly for studies that are complicated and high cost. We now give two examples from the field of the difficulty interpreting results from replication studies in primate cognition.

Experimental field studies

The *orienting asymmetry* paradigm has been used to investigate the lateralization of acoustic processing in the brain. In the orienting asymmetry paradigm, a sound is played from behind a subject

and it is noted whether the subject turns towards the sound with the right ear or the left ear leading. Hauser and Andersson (1994) introduced this paradigm with rhesus monkeys (*Macaca mulatta*) on Cayo Santiago, reporting that adult, but not infant, monkeys showed a right ear bias for conspecific calls, and a left ear bias for heterospecific calls. Because sounds from primate ears are processed in the contralateral hemisphere of the brain, Hauser and Andersson claimed that macaques process species-specific calls in the left hemisphere and heterospecific calls in the right hemisphere. The orienting asymmetry paradigm therefore offered a simple method to test straightforward, but exciting, hypotheses about animal cognition.

A suite of conceptual replication studies followed, testing various lateralization hypotheses across a variety of species. However, the clean results of Hauser and Andersson were never recovered; Teufel et al. (2007) found no turning biases in Barbary macaques, Scheumann and Zimmermann (2008) found weak evidence for left turning biases for heterospecific calls in male mouse lemurs, but not in females, and Gil-da-Costa and Hauser (2006) found the *opposite* pattern in captive vervet monkeys — a left turning bias for species-specific calls. Interpreting these data proved difficult. On the one hand, auxiliary hypotheses could be built around theories to let them fully explain the data. For example, the lateralized responses could *really* be stronger in male mouse lemurs than female mouse lemurs and vervet monkeys could *really* have evolved the opposite lateralization pattern to other primates. However, such attempts to retrofit theories to the data are often weak, and only possible with highly flexible and near unfalsifiable theories (Lakatos, 1970; Roberts & Pashler, 2000). On the other hand, the results could be better explained by a mixture of true and false positive results, confounded designs, low power research and publication bias, and produce real uncertainty about the status of the underlying theory (Teufel et al. 2010). Such a critique is likely strong, and highlights a problem faced by many other disciplines too: if replications will likely produce conflicting results, how can we synthesize these data appropriately (Stegenga, 2009, 2011)?

Observational field studies

A second example concerns Cheney et al.'s (1995) observational study of chacma baboons' signaling behavior. when approaching subordinate conspecifics. Cheney et al. noted that focal individuals who grunt when approaching conspecifics were less likely to be involved in aggressive interactions with this conspecific and more likely to be involved in affiliative interactions, than individuals who approach silently. Later, Silk et al. (2018) conducted similar observations with olive baboons, also finding that focal individuals who grunted when approaching conspecifics were less likely to be aggressive, and more likely to be affiliative. Faraut and colleagues (2019) then used the same approach in a third species of baboons (Guinea baboons) and once again found that affiliative interactions were also more likely after interactions with grunts, but that aggression was not less

likely. The authors noted that this negative statistical result was likely due to a floor effect, as rates of overt aggression in this species are low.

How should we interpret this body of research? The data consistently suggest that baboon grunts signal benign intent, and relative to this claim the Silk et al. (2018) and Faraut et al. (2019) studies can be considered replications of Cheney et al. (1995), despite not being labelled as such (Machery, 2020; Nosek & Errington, 2020). However, our certainty in this claim should be modulated by a quality and bias assessment of the previous studies: how strong a test of the claim was each study, and how likely is it that the body of research has been affected by publication bias, or spin? If we suspect low, then we should be confident in the overall claim, without needing any direct replication studies.

Rethinking Replication: A Sampling Definition

Recently, sampling definitions of replication have been developed in order to better understand the role of replication in science (Asendorpf et al., 2013; Machery, 2020), and animal research (Farrar et al., 2020b; Halina, 2020). Sampling approaches to replication consider how experiments sample from populations across many different levels — populations of experimental units, settings, treatments and measurements (Machery, 2020). In order to test a theory or claim, experiments attempt to sample from within the populations where the theory or claim is relevant (Nosek & Errington, 2020). For example, when testing the claim that great-apes' eye movements track the beliefs about conspecifics (Krupenye et al., 2016), it is necessary to sample great apes, whereas when testing the Rescorla-Wagner model of associative learning, many different taxa can be used. From the sampling perspective, replication studies are just like any other scientific study — a test of a theory or claim. A replication study attempts to re-sample from the same populations that an original study sampled from, re-testing the same scientific claim. In the case of direct replications, the claims at hand would have narrow populations associated with them, and for more conceptual replications, the claims would have wider populations associated with them.

Designing Strong Replication Studies

A strength of the sampling approach to replication is that it drives researchers to consider how well original and replication studies test certain claims, rather than by arguing over whether the experiments were suitably similar to each other (Farrar et al., 2020b). A strong test of a claim would sample widely throughout the populations of interest, for example testing primates at multiple sites,

with multiple different experiments and many different, validated, stimuli. A weak test of the claim, on the other hand, might test a few individuals of one species at a single site with a single stimulus set (Baribault et al., 2018). For primate cognition research, the same feature that makes individual studies difficult to replicate (i.e. between-site variation) can mean that the overall output of multi-laboratory studies is robust. Given that they have no option but to sample from heterogeneous sties, multi-site studies of primates can effectively detect and avoid false positive results.

Researchers face a trade-off between homogeneous and heterogeneous sampling when designing studies, regardless if they are replications or not. Historically, many scientists have prized homogeneity: tightly controlling experiments in highly standardized conditions. Homogeneity and control facilitate very direct replication by making it easy for independent teams of researchers to closely repeat experiments. However, this process does not guarantee replicable results: in some of the most standardized animal research, pre-clinical studies on rodents in standardized conditions and with known genetics, results can still be difficult to replicate between laboratories (Crabbe et al., 1999; Wahlsten et al., 2003). The debate between heterogeneity and homogeneity can be viewed from many perspectives; internal and external validity (Cartwright, 2007), pseudoreplication (Davies & Gray, 2015; Schank & Koehnle, 2009), or generalizability (Yarkoni, 2019). Ultimately, when testing a claim or theory, researchers must focus on what this claim or theory is, what sources of variation (species, rearing history, context, etc.) may impact the constructs of interest, which populations the theory can be tested in, and attempt to sample from throughout these populations. In replication studies, these populations are derived from previously published claims, which the replication study then re-tests.

Importantly, sampling effectively throughout the populations of interest need not entail sampling tens or hundreds of animals or sites; many questions can be effectively answered by sampling widely from within several individual animals that constitute the population of interest in themselves (Smith & Little, 2018). Increasing sample size may only be necessary if researchers are genuinely interested in estimating aggregated statistics about a group of animals, or if a question necessitates a study design that can only use a single or a few trials.

Incentivizing Replication and Future Directions in Primate Cognition

Understanding whether primate cognition faces a replication crisis akin to other disciplines is an important question for the field, and one that will be facilitated by well-designed direct replication studies. Incentivizing researchers to perform these studies is a first step towards achieving this. Encouragingly, barriers to replication are being deconstructed across scientific bodies. Many journals

now actively encourage discussions and publications of replication studies, and have adopted formats such as registered reports that allow for a results-blind review process (Beran, 2018; Vonk & Krause, 2018). Funding agencies, such as the German Science Foundation, now explicitly offer funding for replication studies (Deutsche Forschungsgemeinschaft, 2017), and along with employers, have signed initiatives such as the Declaration on Research Assessment (DoRA, sfdora.org), which discourages the use of impact factors, amongst other criteria, when assessing research and quality of researchers. As awareness of replication problems in science increases, researchers should be less likely to take findings at face value, and subsequently more likely to perform replication studies before using previous findings as a basis for future research.

However, the replication crisis has had a further impact on science aside from assessing the strength of previously published findings. Notably, research has focused on how to improve the strength of evidence that studies produce. To improve the strength of evidence they generate, studies should use a design to test a claim with high statistical power to detect theoretically meaningful effect sizes, and work to establish the validity of this design. This process could be facilitated by developing and testing theoretical and mathematical models (Allen, 2014; Guest & Martin, 2020; Lee et al., 2019; Lind, 2018; Smith et al., 2012; van Rooij & Baggio, 2020). Such models can help researchers assess the similarity of the statistical model at hand to the theoretical claim of interest (Yarkoni, 2019), as long as this evaluation is rigorous (Roberts & Pashler, 2000). Primate cognition studies vary in the feasibility of developing such models. In research programmes that have large control over environmental variables and involve many hundreds of trials, often with highly trained animals, strong tests of precise theories are possible, as are many close replication studies. In contrast, in research programmes that cannot effectively control environmental variables, and cannot collect a large amount of data from each individual participant, these strong tests are less feasible. Such a divide in the research methods of primate cognition means that replication may have different roles to play in different research programmes. In the highly controlled, many-trial programmes, replication and extension are an essential part of mature theory development and incremental science (Bonett, 2012; Nosek & Errington, 2020). In the less-controlled programmes, replications can help to identify more robust statistical effects, from which many speculative theories can be proposed and discussed.

Summary

Replication studies are a useful tool to both assess the robustness of claims in primate cognition research, and to develop theories of primate cognition. Direct replications may make up less than 10% of published experiments in primate cognition research, and they may be clustered around relatively

simple and easy-to-run experiments. Because of this, it is currently unknown just how much of a problem primate cognition faces regarding the replicability of its research. Through a combination of replication studies and meta-research, primate cognition research will be able to retrospectively assess the strength of its published research findings. Prospectively, replication-based research can help to shift incentives towards greater scientific rigor, transparency and theory development in primate cognition, as well as encourage broader discussions of the overall goal of primate cognition research.

References

Alexander, D. M., & Moors, P. (2018). If we accept that poor replication rates are mainstream. *Behavioral and Brain Sciences, 41,* e121. https://doi.org/10.1017/S0140525X18000572

Allen, C. (2014). Models, mechanisms, and animal minds. *The Southern Journal of Philosophy, 52,* 75–97. https://doi.org/10.1111/sjp.12072

Anderson, J. R., & Gallup, G. G. (2011). Which primates recognize themselves in mirrors? *PLoS Biology, 9*(3), e1001024. https://doi.org/10.1371/journal.pbio.1001024

Asendorpf, J. B., Conner, M., Fruyt, F. D., Houwer, J. D., Denissen, J. J. A., Fiedler, K., Fiedler, S., Funder, D. C., Kliegl, R., Nosek, B. A., Perugini, M., Roberts, B. W., Schmitt, M., Aken, M. A. G. van, Weber, H., & Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*(2), 108–119. https://doi.org/10.1002/per.1919

Baribault, B., Donkin, C., Little, D. R., Trueblood, J. S., Oravecz, Z., van Ravenzwaaij, D., White, C. N., De Boeck, P., & Vandekerckhove, J. (2018). Metastudies for robust tests of theory. *Proceedings of the National Academy of Sciences*, *115*(11), 2607–2612. https://doi.org/10.1073/pnas.1708285114

Barrett, L. (2015). Why brains are not computers, why behaviorism is not satanism, and why dolphins are not aquatic apes. *The Behavior Analyst*, *39*(1), 9–23. https://doi.org/10.1007/s40614-015-0047-0

Beach, F. A. (1950). The snark was a boojum. *American Psychologist*, *5*(4), 115–124. https://doi.org/10.1037/h0056510

Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature, 483,* 531–533. https://doi.org/10.1038/483531a

Beran, M. J. (2001). Summation and numerousness judgments of sequentially presented sets of items by chimpanzees (*Pan troglodytes*). *Journal of Comparative Psychology*, *115*(2), 181–191. https://doi.org/10.1037/0735-7036.115.2.181

Beran, M. J. (2018). *Replication and pre-registration in comparative psychology*. *International Journal of Comparative Psychology, 31*. https://escholarship.org/uc/item/59f4z2nd

Bonett, D. G. (2012). Replication-Extension Studies. *Current Directions in Psychological Science*, *21*(6), 409–412. https://doi.org/10.1177/0963721412459512

Brecht, K. F., Müller, J., & Nieder, A. (2020). Carrion crows (*Corvus corone corone*) fail the mirror mark test yet again. *Journal of Comparative Psychology*, *134*(4), 372–378. https://doi.org/10.1037/com0000231

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., … Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637. https://doi.org/10.1038/s41562-018-0399-z

Candea, M. (2013). Habituating meerkats and redescribing animal behaviour science. *Theory, Culture & Society*, *30*(7–8), 105–128. https://doi.org/10.1177/0263276413501204

Cartwright, N. (2007). Are RCTs the gold standard? *BioSocieties*, *2*(1), 11–20. https://doi.org/10.1017/S1745855207005029

Chapelain, A. S., Hogervorst, E., Mbonzo, P., & Hopkins, W. D. (2011). Hand preferences for bimanual coordination in 77 bonobos (*pan paniscus*): replication and extension. *International Journal of Primatology*, *32*(2), 491–510. https://doi.org/10.1007/s10764-010-9484-5

Cheney, D. L., Seyfarth, R. M., & Silk, J. B. (1995). The role of grunts in reconciling opponents and facilitating interactions among adult female baboons. *Animal Behaviour*, *50*(1), 249–257. https://doi.org/10.1006/anbe.1995.0237

Colombo, M., & Scarf, D. (2020). are there differences in "intelligence" between nonhuman species? the role of contextual variables. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.02072

Crabbe, J. C., Wahlsten, D., & Dudek, B. C. (1999). Genetics of mouse behavior: interactions with laboratory environment. *Science*, *284*(5420), 1670–1672. https://doi.org/10.1126/science.284.5420.1670

Craig, D. P. A., & Abramson, C. I. (2018). Ordinal pattern analysis in comparative psychology—A flexible alternative to null hypothesis significance testing using an observation oriented modeling paradigm. *International Journal of Comparative Psychology*, *31*. https://escholarship.org/uc/item/08w0c08s

Culina, A., van den Berg, I., Evans, S., & Sánchez-Tójar, A. (2020). Low availability of code in ecology: A call for urgent action. *PLOS Biology*, *18*(7), e3000763. https://doi.org/10.1371/journal.pbio.3000763

Davies, G. M., & Gray, A. (2015). Don't let spurious accusations of pseudoreplication limit our ability to learn from natural experiments (and other messy kinds of ecological monitoring). *Ecology and Evolution*, *5*(22), 5295–5304. https://doi.org/10.1002/ece3.1782

Deutsche Forschungsgemeinschaft. (2017). *DFG Statement on the Replicability of Research Results* (Available at Www.Dfg.de/En/Research_funding/Announcements_proposals/2017/Info_wissenschaft_17_18/).

Dunlap, K. (1926). The experimental methods of psychology. In C. Murchison (Ed.), *Psychologies of 1925*. (pp. 331–351). Clark University Press. https://doi.org/10.1037/11020-022

Eaton, T., Hutton, R., Leete, J., Lieb, J., Robeson, A., & Vonk, J. (2018). Bottoms-up! Rejecting top-down human-centered approaches in comparative psychology. *International Journal of Comparative Psychology*, *31*. https://escholarship.org/uc/item/11t5q9wt

Estrada, A., Garber, P. A., Rylands, A. B., Roos, C., Fernandez-Duque, E., Di Fiore, A., Nekaris, K. A.-I., Nijman, V., Heymann, E. W., Lambert, J. E., Rovero, F., Barelli, C., Setchell, J. M., Gillespie, T. R., Mittermeier, R. A., Arregoitia, L. V., de Guinea, M., Gouveia, S., Dobrovolski, R., … Li, B. (2017). Impending extinction crisis of the world's primates: Why primates matter. *Science Advances*, *3*, e1600946. https://doi.org/10.1126/sciadv.1600946

Faraut, L., Siviter, H., Pesco, F. D., & Fischer, J. (2019). How life in a tolerant society affects the usage of grunts: Evidence from female and male Guinea baboons. *Animal Behaviour*, *153*, 83–93. https://doi.org/10.1016/j.anbehav.2019.05.003

Farrar, B. G., Altschul, D. M., Fischer, J., van der Mescht, J., Placi, S., Troisi, C. A., Vernouillet, A., Clayton, N. S., & Ostojic, L. (2020). Trialling meta-research in comparative cognition: claims and statistical inference in animal physical cognition. *Animal Behavior and Cognition*, *7*(3), 419–444. https://doi.org/10.26451/abc.07.03.09.2020

Farrar, B., G., Boeckle, M., & Clayton, N., S. (2020). Replications in comparative cognition: what should we expect and how can we improve? *Animal Behavior and Cognition*, *7*(1), 1–22. https://doi.org/10.26451/abc.07.01.02.2020

Farrar, B., G., Voudouris, K., & Clayton, N. (2020). *Replications, comparisons, sampling and the problem of representativeness in animal behavior and cognition research*. PsyArXiv. https://doi.org/10.31234/osf.io/2vt4k

Fidler, F., & Wilcox, J. (2018). Reproducibility of scientific results. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2018). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2018/entries/scientific-reproducibility/

Fiedler, K., & Prager, J. (2018). The regression trap and other pitfalls of replication science—illustrated by the report of the open science collaboration. *Basic and Applied Social Psychology*, *40*(3), 115–124. https://doi.org/10.1080/01973533.2017.1421953

Field, S. M., Hoekstra, R., Bringmann, L., & Ravenzwaaij, D. van. (2019). When and why to replicate: as easy as 1, 2, 3? *Collabra: Psychology*, *5*(1), 46. https://doi.org/10.1525/collabra.218

Forss, S., Motes-Rodrigo, A., Hrubesch, C., & Tennie, C. (2020). Chimpanzees' (*Pan troglodytes*) problem-solving skills are influenced by housing facility and captive care duration. *PeerJ*, *8*, e10263. https://doi.org/10.7717/peerj.10263

Francis, G. (2014). The frequency of excess success for articles in *Psychological Science*. *Psychonomic Bulletin & Review*, *21*(5), 1180–1187. https://doi.org/10.3758/s13423-014-0601-x

Fraser, H., Barnett, A., Parker, T. H., & Fidler, F. (2020). The role of replication studies in ecology. *Ecology and Evolution*, *10*(12), 5197–5207. https://doi.org/10.1002/ece3.6330

Fraser, H., Parker, T., Nakagawa, S., Barnett, A., & Fidler, F. (2018). Questionable research practices in ecology and evolution. *PLOS ONE*, *13*(7), e0200303. https://doi.org/10.1371/journal.pone.0200303

Gil-da-Costa, R., & Hauser, M. D. (2006). Vervet monkeys and humans show brain asymmetries for processing conspecific vocalizations, but with opposite patterns of laterality. *Proceedings. Biological Sciences*, *273*(1599), 2313–2318. https://doi.org/10.1098/rspb.2006.3580

Guest, O., & Martin, A. E. (2020). *How computational modeling can force theory building in psychological science* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/rybh9

Halina, M. (2020). *Replications in Comparative Psychology* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/sqxah

Hauser, M. D., & Andersson, K. (1994). Left hemisphere dominance for processing vocalizations in adult, but not infant, rhesus monkeys: Field experiments. *Proceedings of the National Academy of Sciences*, *91*(9), 3946–3948. https://doi.org/10.1073/pnas.91.9.3946

Herrmann, E., Call, J., Hernàndez-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. https://doi.org/10.1126/science.1146282

Hopkins, W. D., Wesley, M. J., Izard, M. K., Hook, M., & Schapiro, S. J. (2004). Chimpanzees (*Pan troglodytes*) are predominantly right-handed: replication in three populations of apes. *Behavioral Neuroscience*, *118*(3), 659–663. https://doi.org/10.1037/0735-7044.118.3.659

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Ioannidis, J. P. A. (2012). Scientific inbreeding and same-team replication: Type D personality as an example. *Journal of Psychosomatic Research*, *73*(6), 408–410. https://doi.org/10.1016/j.jpsychores.2012.09.014

Krasheninnikova, A., Berardi, R., Lind, M.-A., O'Neill, L., & von Bayern, A. M. P. (2019). Primate cognition test battery in parrots. *Behaviour*, *156*(5–8), 721–761. https://doi.org/10.1163/1568539X-0003549

Krasheninnikova, A., Chow, P. K. Y., & von Bayern, A. M. P. (2020). Comparative cognition: Practical shortcomings and some potential ways forward. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, *74*(3), 160–169. https://doi.org/10.1037/cep0000204

Krupenye, C., Kano, F., Hirata, S., Call, J., & Tomasello, M. (2016). Great apes anticipate that other individuals will act according to false beliefs. *Science*, *354*(6308), 110–114. https://doi.org/10.1126/science.aaf8110

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the Growth of Knowledge* (pp. 91–196). Cambridge University Press. https://doi.org/10.1017/CBO9781139171434.009

Latour, B., & Woolgar, S. (1986). *Laboratory life: The construction of scientific facts*. Princeton University Press.

Lee, M. D., Criss, A. H., Devezer, B., Donkin, C., Etz, A., Leite, F. P., Matzke, D., Rouder, J. N., Trueblood, J. S., White, C. N., & Vandekerckhove, J. (2019). Robust Modeling in Cognitive Science. *Computational Brain & Behavior*, *2*(3–4), 141–153. https://doi.org/10.1007/s42113-019-00029-y

Leonelli, S. (2018, January 28). *Re-Thinking Reproducibility as a Criterion for Research Quality* [Preprint]. http://philsci-archive.pitt.edu/14352/

Lind, J. (2018). What can associative learning do for planning? *Royal Society Open Science*, *5*(11), 180778. https://doi.org/10.1098/rsos.180778

Llorente, M., Riba, D., Palou, L., Carrasco, L., Mosquera, M., Colell, M., & Feliu, O. (2011). Population-level right-handedness for a coordinated bimanual task in naturalistic housed chimpanzees: Replication and extension in 114 animals from Zambia and Spain. *American Journal of Primatology*, *73*(3), 281–290. https://doi.org/10.1002/ajp.20895

Machery, E. (2020). What is a replication? *Philosophy of Science*, 709701. https://doi.org/10.1086/709701

MacLean, E. L., Hare, B., Nunn, C. L., Addessi, E., Amici, F., Anderson, R. C., Aureli, F., Baker, J. M., Bania, A. E., Barnard, A. M., Boogert, N. J., Brannon, E. M., Bray, E. E., Bray, J., Brent, L. J. N., Burkart, J. M., Call, J., Cantlon, J. F., Cheke, L. G., … Zhao, Y. (2014). The evolution of self-control. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(20), E2140-2148. https://doi.org/10.1073/pnas.1323533111

Many Primates, Altschul, D., Beran, M. J., Bohn, M., Caspar, K., Fichtel, C., Försterling, M., Grebe, N., Hernandez-Aguilar, R. A., Kwok, S. C., Rodrigo, A. M., Proctor, D., Sanchez-Amaro, A., Simpson, E. A., Szabelska, A., Taylor, D., van der Mescht, J., Völter, C., & Watzek, J. (2019). *Collaborative open science as a way to reproducibility and new insights in primate cognition research* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/8w7zd

Many Primates, Altschul, D. M., Beran, M. J., Bohn, M., Call, J., DeTroy, S., Duguid, S. J., Egelkamp, C. L., Fichtel, C., Fischer, J., Flessert, M., Hanus, D., Haun, D. B. M., Haux, L. M., Hernandez-Aguilar, R. A., Herrmann, E., Hopper, L. M., Joly, M., Kano, F., … Watzek, J. (2019). Establishing an infrastructure for collaboration in primate cognition research. *PLOS ONE*, *14*(10), e0223675. https://doi.org/10.1371/journal.pone.0223675

Mendes, N., Rakoczy, H., & Call, J. (2011). Primates do not spontaneously use shape properties for object individuation: A competence or a performance problem? *Animal Cognition*, *14*(3), 407–414. https://doi.org/10.1007/s10071-010-0375-0

Minocher, R., Atmaca, S., Bavero, C., McElreath, R., & Beheim, B. (2020). *Reproducibility improves exponentially over 63 years of social learning research*. PsyArXiv. https://doi.org/10.31234/osf.io/4nzc7

Motes Rodrigo, A., Ramirez Torres, C. E., Hernandez Salazar, L. T., & Laska, M. (2018). Hand preferences in two unimanual and two bimanual coordinated tasks in the black-handed spider monkey (*Ateles geoffroyi*). *Journal of Comparative Psychology*, *132*(2), 220–229. https://doi.org/10.1037/com0000110

Murphy, T., Dias, G. P., & Thuret, S. (2014). Effects of diet on brain plasticity in animal and human studies: mind the gap *Neural Plasticit*y 2014. https://doi.org/10.1155/2014/563160

Nelson, E. L., Figueroa, A., Albright, S. N., & Gonzalez, M. F. (2015). Evaluating handedness measures in spider monkeys. *Animal Cognition*, *18*(1), 345–353. https://doi.org/10.1007/s10071-014-0805-5

Neuliep, J. W., & Crandall, R. (1990). Editorial bias against replication research. *Journal of Social Behavior & Personality*, *5*(4), 85–90.

Nissen, S. B., Magidson, T., Gross, K., & Bergstrom, C. T. (2016). Publication bias and the canonization of false facts. *ELife*, *5*, e21451. https://doi.org/10.7554/eLife.21451

Nosek, B. A., & Errington, T. M. (2020). What is replication? *PLOS Biology*, *18*(3), e3000691. https://doi.org/10.1371/journal.pbio.3000691

Open Science Collaboration, O. S. (2015). PSYCHOLOGY. Estimating the reproducibility of psychological science. *Science (New York, N.Y.)*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Palmer, A. R. (2000). Quasi-replication and the contract of error: lessons from sex ratios, heritabilities and fluctuating asymmetry. *Annual Review of Ecology and Systematics*, *31*(1), 441–480. https://doi.org/10.1146/annurev.ecolsys.31.1.441

Peterson, D. (2016). The baby factory: difficult research objects, disciplinary standards, and the production of statistical significance. *Socius*, *2*, 2378023115625071. https://doi.org/10.1177/2378023115625071

Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery*, *10*(9), 712–712. https://doi.org/10.1038/nrd3439-c1

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*(2), 358–367.

Rowe, C., & Healy, S. D. (2014). Measuring variation in cognition. *Behavioral Ecology*, *25*(6), 1287–1292. https://doi.org/10.1093/beheco/aru090

Schank, J. C., & Koehnle, T. J. (2009). Pseudoreplication is a pseudoproblem. *Journal of Comparative Psychology* , *123*(4), 421–433. https://doi.org/10.1037/a0013579

Schauer, J. M., & Hedges, L. V. (2020a). Reconsidering statistical methods for assessing replication. *Psychological Methods*. https://doi.org/10.1037/met0000302

Schauer, J. M., & Hedges, L. V. (2020b). Assessing heterogeneity and power in replications of psychological experiments. *Psychological Bulletin*, *146*(8), 701–719. https://doi.org/10.1037/bul0000232

Scheel, A. M., Schijen, M., & Lakens, D. (2020). *An excess of positive results: Comparing the standard Psychology literature with Registered Reports* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/p6e9c

Scheel, M. H., Shaw, H. L., & Gardner, R. A. (2016). Incomparable methods vitiate cross-species comparisons: a comment on Haun, Rekers, and Tomasello (2014). *Psychological Science*, *27*(12), 1667–1669. https://doi.org/10.1177/0956797615595229

Scheumann, M., & Zimmermann, E. (2008). Sex-specific asymmetries in communication sound perception are not related to hand preference in an early primate. *BMC Biology*, *6*(1), 3. https://doi.org/10.1186/1741-7007-6-3

Schmitt, V., Pankau, B., & Fischer, J. (2012). Old World Monkeys Compare to Apes in the Primate Cognition Test Battery. *PLoS ONE*, *7*(4), e32024. https://doi.org/10.1371/journal.pone.0032024

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. https://doi.org/10.3389/fpsyg.2020.01835

Schubiger, M. N., Kissling, A., & Burkart, J. M. (2016). How task format affects cognitive performance: A memory test with two species of New World monkeys. *Animal Behaviour*, *121*, 33–39. https://doi.org/10.1016/j.anbehav.2016.08.005

Silk, J. B., Roberts, E. R., Städele, V., & Strum, S. C. (2018). To grunt or not to grunt: Factors governing call production in female olive baboons, Papio anubis. *PLOS ONE*, *13*(11), e0204601. https://doi.org/10.1371/journal.pone.0204601

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*(2), 534–547. https://doi.org/10.1037/a0033242

Skinner, B. F. (1956). A case history in scientific method. *American Psychologist*, *11*(5), 221–233. https://doi.org/10.1037/h0047662

Smith, J. D., Couchman, J. J., & Beran, M. J. (2012). The highs and lows of theoretical interpretation in animal-metacognition research. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1297–1309. https://doi.org/10.1098/rstb.2011.0366

Smith, P. L., & Little, D. R. (2018). Small is beautiful: In defense of the small-N design. *Psychonomic Bulletin & Review*, *25*(6), 2083–2101. https://doi.org/10.3758/s13423-018-1451-8

Stegenga, J. (2009). Robustness, Discordance, and Relevance. *Philosophy of Science*, *76*(5), 650–661. JSTOR. https://doi.org/10.1086/605819

Stegenga, J. (2011). Is meta-analysis the platinum standard of evidence? *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *42*(4), 497–507. https://doi.org/10.1016/j.shpsc.2011.07.003

Teufel, C., Hammerschmidt, K., & Fischer, J. (2007). Lack of orienting asymmetries in Barbary macaques: Implications for studies of lateralized auditory processing. *Animal Behaviour*, *73*(2), 249–255. https://doi.org/10.1016/j.anbehav.2006.04.011

van Rooij, I., & Baggio, G. (2020). *Theory before the test: How to build high-verisimilitude explanatory theories in psychological science*. PsyArXiv. https://doi.org/10.31234/osf.io/7qbpr

Vonk, J., & Krause, M. (2018). Editorial: Announcing preregistered reports. *Animal Behavior and Cognition*, *5*(2), i–ii. https://doi.org/10.26451/abc.05.02.00.2018

Wahlsten, D., Metten, P., Phillips, T. J., Boehm, S. L., Burkhart-Kasch, S., Dorow, J., Doerksen, S., Downing, C., Fogarty, J., Rodd-Henricks, K., Hen, R., McKinnon, C. S., Merrill, C. M., Nolte, C., Schalomon, M., Schlumbohm, J. P., Sibert, J. R., Wenger, C. D., Dudek, B. C., & Crabbe, J. C. (2003). Different data from different labs: Lessons from studies of gene-environment interaction. *Journal of Neurobiology*, *54*(1), 283–311. https://doi.org/10.1002/neu.10173

Wallace, E. K., Altschul, D., Körfer, K., Benti, B., Kaeser, A., Lambeth, S., Waller, B. M., & Slocombe, K. E. (2017). Is music enriching for group-housed captive chimpanzees (*Pan troglodytes*)? *PLOS ONE*, *12*(3), e0172672. https://doi.org/10.1371/journal.pone.0172672

Yarkoni, T. (2019). *The generalizability crisis* . PsyArXiv. https://doi.org/10.31234/osf.io/jqw35

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, *41*, e120. https://doi.org/10.1017/S0140525X17001972