ABC 2021, 8(2):273-295
Animal Behavior and Cognition
DOI: https://doi.org/10.26451/abc.08.02.14.2021

# Replications, Comparisons, Sampling and the Problem of Representativeness in Animal Cognition Research

**Benjamin G. Farrar*, Konstantinos Voudouris, & Nicola S. Clayton**

Department of Psychology, University of Cambridge

*Corresponding author (Email: bgf22@cam.ac.uk)

**Citation –** Farrar, B. G., Voudouris, K., & Clayton, N. S. (2021). Replications, comparisons, sampling and the problem of representativeness in animal cognition research. *Animal Behavior and Cognition*, *8*(*2*), 273-295. https://doi.org/10.26451/abc.08.02.14.2021

**Abstract –** Animal cognition research often involves small and idiosyncratic samples. This can constrain the generalizability and replicability of a study's results and prevent meaningful comparisons between samples. However, there is little consensus about what makes a strong replication or comparison in animal research. We apply a resampling definition of replication to answer these questions in Part 1 of this article, and, in Part 2, we focus on the problem of representativeness in animal research. Through a case study and a simulation study, we highlight how and when representativeness may be an issue in animal behavior and cognition research and show how the representativeness problems can be viewed through the lenses of, i) replicability, ii) generalizability and external validity, iii) pseudoreplication and, iv) theory testing. Next, we discuss when and how researchers can improve their ability to learn from small sample research through, i) increasing heterogeneity in experimental design, ii) increasing homogeneity in experimental design, and, iii) statistically modeling variation. Finally, we describe how the strongest solutions will vary depending on the goals and resources of individual research programs and discuss some barriers towards implementing them.

**Keywords –** Animal cognition, Comparison, Experimental design, Generalizability, Replication, Representativeness, Sampling

Animal cognition research often involves small samples. In order to make general claims about a group or species' behavior, researchers assume that their samples are representative enough of the group or species of interest. However, this assumption is rarely tested, and the literature is populated by claims that are produced by single laboratories, testing the same animals, at single time points and in closely related experimental designs. This could lead to overgeneralized findings that are difficult to replicate (Baker, 2016; Henrich et al., 2010; Würbel, 2000; Yarkoni, 2019), but equally, it could be an effective strategy to maximize scientific progress in resource-limited fields (Craig & Abramson, 2018; Davies & Gray, 2015; Mook, 1983; Schank & Koehnle, 2009; Smith & Little, 2018). To explore this issue, this article shows how concerns about replicability, representativeness, comparison and theory testing, and pseudoreplication are all related through the lens of sampling. To design the best experiments, researchers should consider all five in relation to their sampling plans. Part 1 of this article focuses on sampling and replication, and answers the following questions:

- What is a replication in animal behavior and cognition research?
- What is the relationship between replication and theory testing?
- What makes a species-fair comparison?

Part 2 of the article then focuses on representativeness and asks how concerned researchers should be with the problem of non-representative sampling in animal research. We explore this issue through a re-analysis of existing data on animal 'self-control' and a simulation study. The simulation study shows that, for some between-group or between-species comparisons, poorly representative samples could lead to false positive rates closer to 50% than 5%, the rate conventionally cited when authors use $p < .05$ to define statistical significance. Finally, we end the article with a discussion of how researchers might assess, mitigate, and account for the problem of representativeness in comparative cognition.

## Part 1 – Claims, Samples and Replications

### What Are Replications in Animal Research?

A study is labeled a replication because it is the same in some regards to a previous experiment. For example, a replication study may repeat the same experimental protocol as a previous study, except use a new sample of animals. However, it is not possible to perform *exactly* the same study twice, and because of this, any replication study can also be reframed in terms of a test of generalization. Even if the same experimenters perform the same experiment on the same group of animals, the replication experiment is still a test of generalization across time.

However, while truly identical replications are impossible, this does not mean the concept of replication is obsolete, or redundant with generalizability. When performing replications, scientists are not usually interested in what philosophers call absolute identity, but in what they call relative identity (Geach, 1973; Lewis, 1993; Noonan & Curtis, 2004; Quine, 1950). They are not interested in whether a feature of a replication is exactly the same as an original study, rather, they are interested in whether that feature *can be considered* the same, or as coming from the same population, relative to a given theory. Idealistically, a theory or claim would specify what can and cannot be considered as coming from the same population, i.e., identifying its boundary conditions (e.g., Simons et al., 2017), and thus what a valid test of it would sample from. For example, consider the Rescorla-Wagner model, which specifies that gains in associative strength are proportional to the prediction error (Rescorla & Wagner, 1972). From the perspective of the Rescorla-Wagner model, it does not matter whether the hypothesis is tested with a sample of rats or a sample of mice, or pigeons, or monkeys, etc. Providing a valid conditioning procedure is followed, all of these species are within the boundary conditions of the Rescorla-Wagner model, and an original study making a general claim about the Rescorla-Wagner model by testing rats could therefore be replicated in pigeons or in monkeys – the Rescorla-Wagner model makes no distinction. On the contrary, the most robust tests of the Rescorla-Wagner model would sample from across all of species that the model applies to, rather than just a single species.

Recently, resampling definitions of replication have been developed (Asendorpf et al., 2013; Machery, 2020). These may be the most effective definition of replication in animal cognition research. When researchers test a claim, they sample from populations of experimental units (most often animals), settings, treatments, and measurements (Gómez et al., 2010). For example, when testing the claim that chimpanzees will explore a mark on their forehead when exposed to a mirror, researchers sample from the population of chimpanzees available for research, from various settings (laboratories, zoos, wild), with a variety of possible treatments (different size mirrors, different types of marks, etc.), and many different possible measurements (e.g., an ethogram of self-directed actions). The resampling definition of replication states that a replication study is a study that resamples from the same populations of experimental units, treatments, measurements, and settings that an original study could have sampled from, relative to the claim being tested (Machery, 2020; Nosek & Errington, 2020). This is outlined in Table 1, which is adapted from Machery (2020).

According to the resampling approach, a complete replication resamples from the same populations of experimental units, treatments, measurements, and settings as an original study, relative to the theory or claim in question. However, an experiment could also replicate some features of an original study but not others (Machery, 2020). This would create an explicit test of generalizability; probing whether the claim or theory can be applied successfully outside of some of its pre-specified boundary conditions. For example, a researcher would be able to test whether theories built on work with captive monkeys generalize to their wild counterparts by resampling from the *same* population of, treatments and measurements, but sampling from a *different* population of settings (captive versus wild).

**Table 1**

*A Resampling Account of Replication (Adapted from Machery, 2020)*

| An experiment samples from: | A replication *resamples* from: |
|---|---|
| A population of experimental units,<br>e.g., a population of a species in captivity | The same population of experimental units |
| A population of treatments,<br>e.g., experimental conditions | The same population of treatments |
| A population of measurements,<br>e.g., definitions of success on a trial | The same population of measurements |
| A population of settings,<br>e.g., sites and times | The same population of settings |

To see how the resampling definition can be applied in animal cognition research, we now discuss a partial or "conceptual" replication of a study investigating aging in monkeys (Almeling et al., 2016; Bliss-Moreau & Baxter, 2019). This is a useful example as, like most experiments in animal cognition, Bliss-Moreau and Baxter's study is not a close replication of the previous study; it was neither conducted in identical laboratory settings nor even in the same model species.

## Case Study: Do Nonhuman Primates Lose Interest in the Non-Social World with Age?

In 2016, Almeling and colleagues examined the relationship between the age of monkeys and their interest in the social and non-social environment. They tested 116 Barbary macaques housed in a large (20 ha) outdoor park in France. Across three non-social novel object interest tasks, Almeling et al. reported that older Barbary macaques interacted less with objects compared to younger Barbary macaques ($N$ = 88 in these tasks). From this, they made the general claim that nonhuman primates lose interest in the non-social world with age. Bliss-Moreau and Baxter (2019) replicated one of the object conditions of Almeling et al. in a larger sample of 243 rhesus macaques. However, these rhesus macaques were housed in indoor cages either alone or with a social pair mate, in contrast to the free-roaming Barbary macaques. Bliss-Moreau and Baxter labeled their study as a "conceptual" replication because they tested a different species in a markedly different setting and used a different, albeit conceptually similar, food-baited apparatus. However, relative to the claim that monkeys, in general, display a loss of interest to non-social stimuli with age, the populations sampled by Bliss-Moreau and Baxter do seem to come from the same overall populations that Almeling et al.'s claim specifies, i.e., both are tests of the claim that interest in the non-social environment declines during aging in monkeys.

Bliss-Moreau and Baxter reported no statistically significant effect of age on exploration across the first two minutes, which they interpreted as contrary to the results of Almeling et al. (2016) and challenging "the notion that interest in the 'non-social world' declines with age in macaque monkeys, generally" (Bliss-Moreau & Baxter, 2019, p. 6). This claim seems reasonable: both Almeling et al. and Bliss-Moreau and Baxter sampled from within the experimental units, setting, treatment, and measurement populations implicitly specified by the claim that interest in the non-social world declines with age in macaque monkeys, and so our confidence in the claim overall should decrease, following the negative replication results. But can we really say that Bliss-Moreau and Baxter's experiment *replicated* Almeling et al.'s? This question is difficult, because replications exist on many levels (across experimental units, settings, treatments, and measurements) and are theory or claim dependent. Moreover, most experiments in animal behaviour and cognition do not make a single isolated claim. For example, the following theoretical claims could reasonably be inferred from the Almeling et al. paper:

1) Socially living Barbary macaques lose interest in the non-social environment with age
2) Barbary macaques lose interest in the non-social environment with age
3) Socially living monkeys lose interest in the non-social environment with age
4) Monkeys lose interest in the non-social environment with age

When asking how Bliss-Moreau and Baxter's (2019) study is a replication of Almeling et al.'s, (2016) we should consider not just how the studies relate to each other, but how they relate to each claim we are assessing. Ultimately, the goal of a replication study is usually to test a scientific claim, rather than just to match a previous study's methods (Nosek & Errington, 2020). Therefore, when interpreting the results of replication studies, researchers should focus on how relevant and diagnostic the data from each study are to the claim(s) in question, rather than just how similar they are. The main strength of the resampling definition of replication — that a replication study resamples from the same populations that an original study could have sampled from, relative to the claim being tested — is that it forces researchers analyzing replication studies to consider exactly what is being tested and how effective the test is, rather than focusing unnecessarily on absolute similarity.

One barrier to identifying and testing claims is that many theories and claims in animal cognition are verbal and vague. This makes it difficult to derive risky predictions of the theories, because their vagueness affords them the flexibility to accommodate nearly any result (Roberts & Pashler, 2000). This could be remedied by formally modelling theories and hypotheses (Farrell & Lewandowsky, 2010; Guest & Martin, 2020), and some suggest these models are key to making progress in understanding animal minds (Allen, 2014), or in understanding what comparative cognition can achieve as a science (Farrar & Ostojić, 2019). These models can be informed by known mechanisms driving animal behavior, such as associative learning (Heyes & Dickinson, 1990; Lind, 2018; Lind et al., 2019), but these need not be preferred to, or even contradict, non-associative models (Bausman & Halina, 2018; Mercado, 2016; Smith et al., 2016). Just like any other scientific tool, formal models need critique from a variety of perspectives; but the benefit of these models is that they facilitate such critique, in comparison with verbal theories that can avoid it.

## Species-Fair Comparisons

The resampling account not only offers a theoretical framework for replications, generalizations, and theory testing in animal cognition research, but it also offers a framework for analyzing between-species comparisons. Between-species comparisons are just tests of the generalizability of an effect across species, and like any other test of generalization, they can be reframed in terms of replication, too. Comparing an effect between a group of chimpanzees and a group of bonobos is the same as testing if the effect generalizes from chimpanzees to bonobos, or replicating a study from chimpanzees in bonobos, and both of these are entailed by a coarser study of whether great apes (chimpanzees, bonobos and orangutans) show the effect in question. Whether the study in question is best described as a comparison, replication or a test of a claim is somewhat moot — it is all three at the same time, relative to claims of different coarseness.

However, there are clearly times when researchers may wish to focus on comparative claims, and this requires sampling from *different* population of experimental units, e.g., different breeds, groups, or species of animals (with the caveat that these could be seen as coming from the same population relative to broader claims). For an ideal comparison between two groups of animals, researchers would sample from different populations of experimental units, and the same populations of treatments, measurements, and settings. Again, “same” here does not mean identical, but the same relative to the claim and experimental unit at hand. For example, consider a researcher who wants to compare the relative response of dolphins (e.g., Hill et al., 2016) to familiar and unfamiliar humans with that of elephants (e.g., Polla et al., 2018). Clearly, the researcher must sample from different populations of experimental units [dolphins, elephants], and a different population of settings [aquatic, non-aquatic]. However, even though the settings are different in absolute terms, they are the same relative to the experimental unit; the dolphins are tested in water, the elephants on land, and this makes the comparison more valid (Clark et al., 2019; Leavens et al., 2019; Tomasello & Call, 2008), or a ‘species-fair’ comparison (Boesch, 2007; Brosnan et al., 2013; Eaton et al., 2018; Tomasello & Call, 2008).

## Part 2 – The Problem of Representativeness in Animal Research

A sampling perspective shines light on why many results in animal research may struggle to replicate. Animal experiments often sample a small number of animals at a single site, using a single apparatus and measurement technique. However, from these small samples come general claims about animal behavior, creating a mismatch between the statistical model and the theoretical claim (Yarkoni, 2019). The statistical model will usually allow generalization to the population that the experimental units were randomly sampled from, for example the population of animals at a given site, (although even then they may not be randomly sampled, see Schubiger et al., 2019), but any inferences to the wider population of interest will be overconfident, unless the population of interest can be justified as the individual animal (Smith & Little, 2018). This is an unavoidable consequence of working with difficult to reach populations (Lange, 2019), but it should be accounted for when building theories. This is important as many aspects of animal behavior vary across samples; for example, due to experimenter effects (Beran, 2012; Bohlen et al., 2014; Cibulski et al., 2014; Pfungst, 2018; Sorge et al., 2014), genetic variation (Fawcett et al., 2014; Johnson et al., 2015; MacLean et al., 2019), housing conditions (Farmer et al., 2019; Hemmer et al., 2019; Würbel, 2001), diets (Davidson et al., 2018; Höttges et al., 2019), and learning/developmental histories (Skinner, 1976).

### Situating the Problem of Representativeness

The problem of representativeness has been discussed from several different angles across scientific literatures, unfortunately poorly connected and with different terminologies. However, they share the similar underlying concern that researchers’ claims are poorly matched by their sampling strategies and statistical models.

#### *Replicability*

First, a lack of representative sampling causes low replicability or results reproducibility (not to be confused with computational reproducibility, e.g., see Culina, van den Berg et al., 2020; Minocher et al., 2020): because of small and non-representative samples of experimental units, settings, treatments, and measurements, sampling variation will mean that laboratories will struggle to replicate or reproduce the results of previous experiments. This argument has featured heavily in rodent phenotyping studies (Crabbe et al., 1999; Kafkafi et al., 2017, 2018; Lewejohann et al., 2006; Richter et al., 2009, 2010, 2011; Wahlsten et al., 2003; Würbel, 2000, 2002).

#### *Generalizability and External Validity*

Second, a lack of representative sampling causes problems of generalizability or external validity: researchers’ claims will not often generalize to novel but related settings (Yarkoni, 2019).

### *Pseudoreplication*

Third, the lack of representative sampling in animal research is usually due to non-random sampling from the population of interest. This leads to pseudoreplication (Hurlbert, 1984; Lazic, 2010) if this non-random sampling is not accounted for in the statistical models, and the consequence is that uncertainty intervals will be overly narrow, and the results will struggle to replicate in new samples – or generalize to them.

### *Theory Testing*

Fourth, the lack of representative sampling produces weak tests of a theory or claim (Baribault et al., 2018): a test that probes only a small sample space of a theory's predictions provides less opportunity for weaknesses in the theory or claim to be found, compared to a test that covers most of the relevant sample space.

## The Difficulty of Identifying Differences Between Groups and Species

That animal behavior differs across space and time makes it difficult to understand whether species or group differences in behavior are really a consequence of real species differences, or whether they are due to the host of other factors that vary between sites. In reality, the observed differences between two groups will be the sum of the real group differences in behavior that are of interest and all other factors that influence animal behavior and vary between sites. When making quantitative between-species and between-group comparisons, they are nearly always confounded by site-specific differences in factors that are not the focus of interest. Lazic (2016) commented on such a scenario in an introductory textbook for laboratory biology: "To make valid inferences, one would need to assume that the effects of [site] are zero. Moreover, as this assumption cannot be checked, the researcher can only hope that [site] effects are absent. Such a design should be avoided" (Lazic, 2016, p. 68).

One may object to this and acknowledge that, while there are many variables that differ between sites but go unmeasured, the net sum of these effects should be close to zero across sites, i.e., they will cancel each other out. However, this would only be the case if there were many variables with small effect that were randomly assigned to each site, and this is not what happens. On the contrary, laboratories or sites differ markedly from each other on a range of variables with large effects (e.g., housing conditions, learning experiences). It is often recognized that animal laboratories are poorly positioned to generate representative data of the species in the wild (Boesch, 2020; Calisi & Bentley, 2009), but what if they are also poorly positioned to generate representative data of the species in laboratories? Taken to the extreme, there may be a laboratory that is testing a sample that is more representative of a species other than its own; for example, a sample of lemurs that have parrot-like self-control, or a sample of hand-reared wolves that behave more like dolphins when presented with a novel object. To highlight the difficulties of making between group or between species inferences across sites, we now consider a case study of between species comparisons made using the cylinder task, and then present a simulation study of how sampling affects comparisons in animal research.

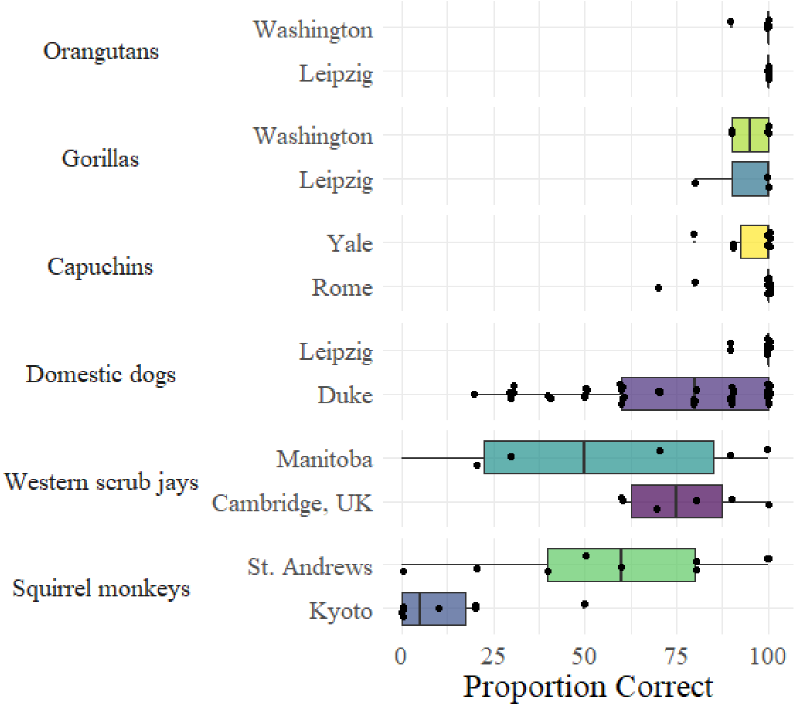
## Case Study: Between Species Comparisons and the Cylinder Task

For this case study, we used data from MacLean et al. (2014) to probe the stability of a measurement of behavioral inhibition when taking new samples of experimental units at new sites. MacLean et al.'s (2014) large-scale study tested the performance of 36 species across 43 sites on two tasks aimed at measuring self-control (but rather measured one form of behavioral inhibition: Beran, 2015); the A not B task and the cylinder task. The cylinder task was given to 32 species across 38 sites. In this task, animals are familiarized with retrieving a piece of food from the center of an opaque cylinder.

After retrieving the food from the opaque cylinder in 4 out of 5 consecutive training trials, the animals proceed to testing. In testing, the animal is presented with a transparent cylinder with food in the center. In order to successfully retrieve the food, the animal needs to inhibit an initial drive to directly reach for the food which would cause them to subsequently collide with the transparent cylinder, and instead detour to the cylinder ends to access the food. Each animal was given 10 trials, and an overall score between 0% (no animals succeeded on any trial) and 100% (all animals succeeded on every trial) was computed for each species. Five species (orangutans, gorillas, capuchin monkeys, squirrel monkeys and domestic dogs) were tested across two sites. Figure 1 displays the between-site variation for these species, and also includes data from an additional species, the Western scrub-jay, that was tested both in the original experiment and a couple of years later at a new site (Stow et al., 2018).

**Figure 1**

*Species Differences Between Sites in the Cylinder Task*



*Note.* All data from MacLean et al. (2014), except the Manitoba scrub-jay data, which are from Stow et al. (2018).

For the species not performing near ceiling (scrub-jays, squirrel monkeys and domestic dogs), the variability is striking. For squirrel monkeys, the median score in Kyoto was 5%, compared with 60% in St Andrews. No individual in Kyoto performed above the median in St Andrews, and this demonstrates how some between-site differences that cannot be attributed to species identity can have large influences on behavior. To highlight the issues this can pose for inference, consider what would happen if the animals from Kyoto were not squirrel monkeys, but Tonkean macaques. Then, it is likely that the difference in performance compared to the St Andrews' squirrel monkeys would likely be interpreted as a species difference – "Tonkean macaques are worse at behavioral inhibition than squirrel monkeys," could be the title of a paper reporting these results. In fact, the substantial difference in behavior between species tested at different sites need not imply meaningful species differences at all. If we took new samples for all species that MacLean et al. tested, it is possible a completely different ranking of animals would be produced. MacLean et al.'s (2014) overall model gains credibility, however, because of the use of phylogenetic models (and also including data from the A-not-B task, another test of behavioral inhibition). Incorporating phylogeny and estimating phylogenetic signal when making comparisons, providing there is enough data, little bias, and sufficient model checks, can lead to large increases in statistical power (Freckleton, 2009; see MacLean et al., 2012 also for an overview of other benefits of comparative phylogenetic models). However, any individual site comparison of non-ceiling cylinder task performance between species, either within the MacLean et al. study, or from other published research, is likely too uncertain to produce meaningful estimates at the species level, and this can lead researchers astray when making inferences from individual results. Table 2 presents some statements from studies that followed MacLean et al.'s procedures using a single species at a single site, along with the species' cylinder task "score."

**Table 2**

*Results and Claims from Four Species Tested on the Cylinder Task*

| Study | Group | Score | Claim |
|---|---|---|---|
| Ferreira et al., 2020 | High ranging chickens | 24% | "High rangers had the worst performance of all species tested thus far" (p. 3) |
| | Low ranging chickens | 40% | |
| Isaksson et al., 2018 | Great tits | 80% | "The average performance of our great tits was 80%, higher than most animals that have been tested and almost in level with the performance in corvids and apes." (p. 1, abstract) |
| Langbein, 2018 | Goats | 63% | "The results indicated that goats showed motor self-regulation at a level comparable to or better than that of many of the bird and mammal species tested to date." (p. 1, abstract) |
| Lucon-Xiccato et al., 2017 | Guppies | 58% | "A performance fully comparable to that observed in most birds and mammals" (p. 1, abstract) |

This set of numerical comparisons are factually correct, but what do they mean? The worst performing chickens actually scored higher than the Kyoto squirrel monkeys, and if we sampled another population of great tits it is possible that their performance would regress close to the mean value of all species. Ordaining a species with a single score following a single test on a small sample of animals from a single site with a single apparatus, and then comparing this number between species has no means of error control and hides the uncertainty in their estimates. Several of the inferences are reasonable; for example, we may genuinely believe that chickens will perform poorly on behavioral inhibition tasks, but this is primarily constrained by our (arbitrary) prior beliefs. For potentially more surprising results, such as the high score of great tits, our beliefs are not so constraining, yet neither are the data.

Moreover, and counter-intuitively, the best estimate of great tit performance on the cylinder task is not the 80% reported by Isaksson et al. (2018), even though this is the only known data collected with great tits on this task. Rather, the best estimate would utilize the information we have about similar animals (other birds of a similar size/socio-ecology/phylogeny), that would shrink our estimate of great tit performance closer to the mean value for, as an example, all Passeriformes tested to date. Interestingly, during the revision process of this article, two further datasets of great tit performance on the cylinder task became available. In contrast to the 80% reported by Isaksson, and in line with our prediction of regression to the mean, Troisi et al. (2020) recorded a score of 38%, and a sample of 35 tested by Coomes et al. (2020) scored 41%. Moreover, in a pilot to one of these studies using a larger tube, a sample of great tits scored 0%, suggesting that the size of the tube can heavily modulate individual's performance (G. L. Davidson, personal communication, Jan., 2021).

How, then, can we make better inferences from single site samples of data? We could attempt to get a better estimate at this single site; for example, by testing great tits on a wide range of tube apparatuses. Alternatively, we can also use the data from other species to inform our great tit estimate. Because the behaviour of different animals will often be correlated, for example as a function of phylogenetic distance, we should allow data from similar species to guide each other's estimates. Ideally, a phylogenetic model would be constructed that incorporates information on the phylogenetic distance between species and a model of the trait's evolution (McElreath, 2016). Other relevant predictor variables, such as body size, tube size, or body size/tube size ratio, could be added into these models, also, or they could be investigated in separate meta-regression models. However, for many animal cognition questions, such models will be difficult to generate, but the general principle holds: when a surprisingly high or surprisingly low estimate of a species behavior is produced, and most data from similar species are less extreme, it is likely that the new estimate is over- or underestimated. Returning to the cylinder task, it is clear that non-ceiling results are not very informative about animal cognition if we do not know whether the results from any given sample are stable across space or time — before considering issues of construct validity (Beran, 2015; Kabadayi et al., 2017, 2018).

## Simulation Study

To illustrate how between-site variation (a proxy for the sum of setting, treatment and measurement variation) can lead to elevated false positive rates and results that struggle to replicate, we now present a short simulation study of a replication and a comparison in comparative cognition. The simulation and visualizations were performed in R 4.0.2 (R Core Team, 2020), using the packages *tidyverse* (Wickham et al., 2019), *extrafont* 0.17 (Chang, 2014) and *scales* 1.1.1 (Wickham & Seidel, 2020). The code is available at: https://github.com/BGFarrar/Replications-Comparisons-and-Sampling. This section can be skipped if the reader is already comfortable with the topic. We simulated a hypothetical within-species replication between two groups of chimpanzees, and a hypothetical between-species replication/comparisons between a group of chimpanzees and a group of bonobos. We simulated 100 hypothetical sites of chimpanzees, and 100 hypothetical sites of bonobos, with 100 animals at each site. The behavior of animals within a site was correlated, such that animals sampled from the same sites, on average, had more similar behaviors than animals sampled from different sites. At each site, we 'measured' each animal's behavior to produce a neophobia and self-control score for each. For both the replication simulation and the comparison simulation, four parameters were used to simulate each animal's behavior: a population grand mean, $\beta_0$, a by-location random intercept $L_0$, a by-subject random intercept $S_0$, and a by-individual residual error term $e_{ls}$. Subject was nested within location, such that all subjects at the same location had the same location effect. Data were simulated using the following formula:

$$Score_{ls} = \beta_0 + L_{0l} + S_{0s} + e_{ls}$$

For the replication simulation, 10,000 chimpanzees were simulated with the following settings:

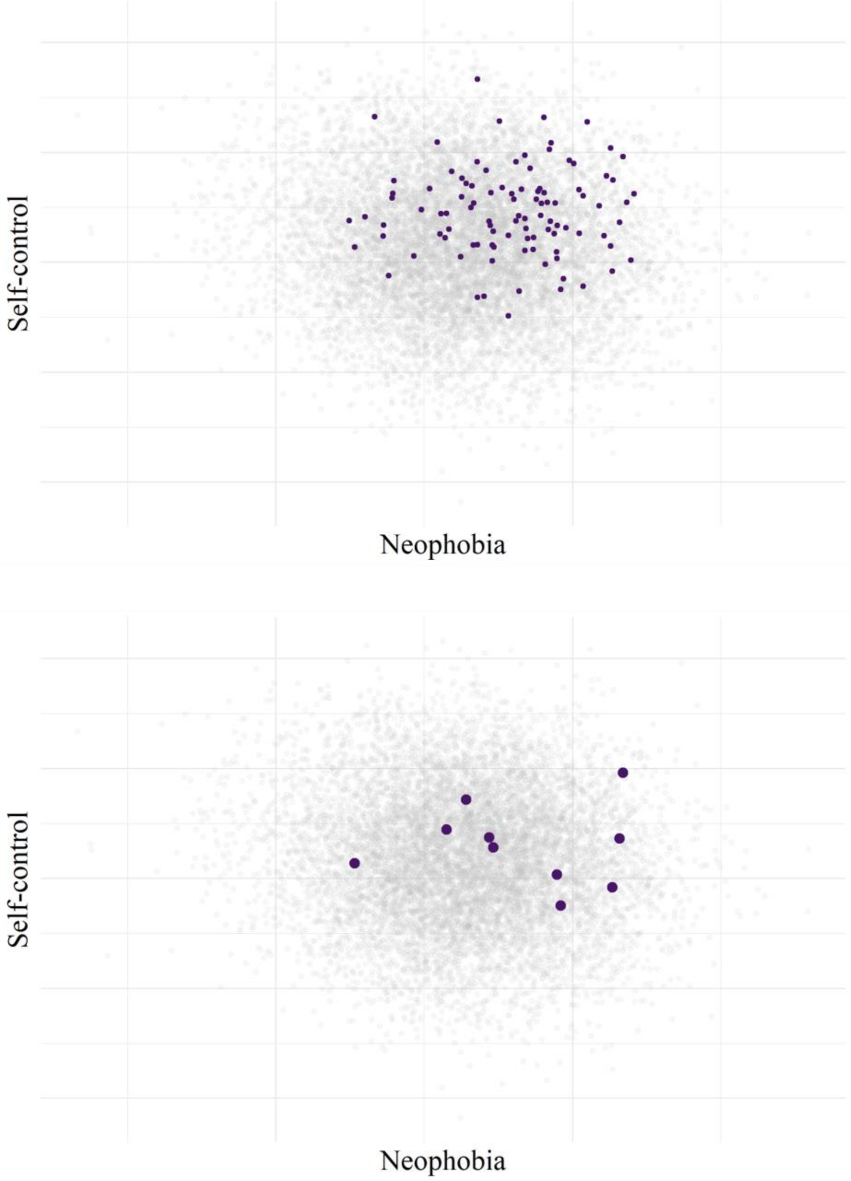
*Neophobia*
$\beta_0$ = 800
$L_{0l}$ ~ $N$(0, 100)
$S_{0s}$ ~ $N$(0, 100)
$e_{ls}$ ~ $N$(0, 50)

*Self-control*
$\beta_0$ = 80
$L_{0l}$ ~ $N$(0, 10)
$S_{0s}$ ~ $N$(0, 10)
$e_{ls}$ ~ $N$(0, 5)

The panels of Figure 2 display the behavior of all 10,000 chimpanzees (100 animals x 100 sites) in grey. Next, we randomly selected one site to be our first sample. The upper panel of Figure 2 highlights all 100 chimpanzees from this site. However, in reality we would not usually have access to or test 100 animals at a site; instead, a primate cognition sample size is usually around 7 (Many Primates, Altschul, Beran, Bohn, Caspar et al., 2019). Therefore, we randomly selected 10 animals, which are highlighted in the lower panel of Figure 2.

**Figure 2**

*The Behavior-Space of a Simulated Population of 10,000 Chimpanzees (grey dots in both panels).*

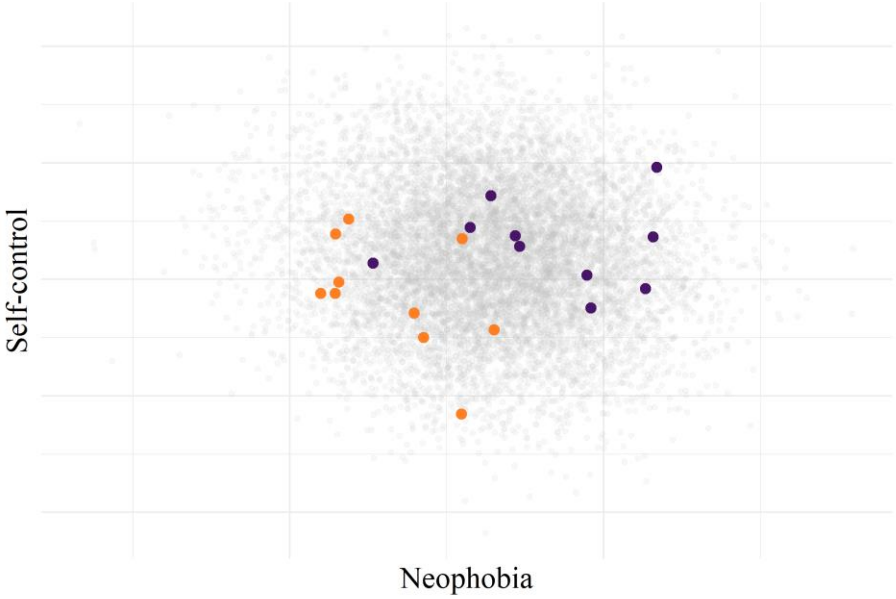


*Note.* In purple, the Upper Panel shows 100 hypothetical chimpanzees sampled from a single site, and the Lower Panel shows just 10 of these chimpanzees.

To create a replication study, we repeated this process, taking another random sample of 10 chimpanzees from a different site. This sample is plotted in Figure 3 alongside our first sample, creating a within-species (or experimental unit) replication, which could also be framed as a between site comparison, or a test of generalizability across sites.

**Figure 3**

*A Hypothetical Within-Species Replication, or Between-Site Comparison*



*Note.* Purple points represent the same chimpanzees sampled from the first site (Figure 2), and orange points represent a second sample of chimpanzees.

The second sample of chimpanzees, in orange, had smaller neophobia and larger self-control scores than the first sample, in purple. Performing a two-sided Welch's *t* test, both differences were statistically significant, $p_{\text{neophobia}} < .001$ and $p_{\text{self-control}} = .04$. This reflects the real variation between the sites, which were simulated at 28% for neophobia, and 14% for self-control. Our samples of just 10 animals captured this difference relatively accurately, estimating the group differences as 31% for neophobia and 14% for self-control. While our two samples provided good estimates of the true between-sample differences, our samples were poorly representative of the overall population of chimpanzees. Site 1 (purple), overestimated neophobia by 14% and self-control by 3%, whereas Site 2 underestimated neophobia by 17%, and self-control by 11%.

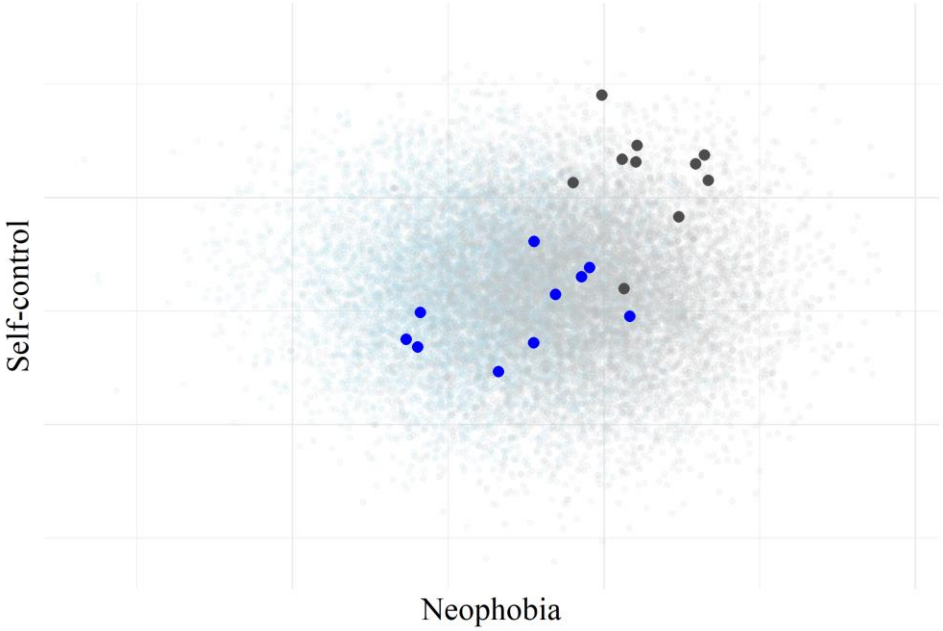
Having simulated a within-species replication, we proceeded to simulate a typical between-species comparison. To achieve this, we randomly sampled from the set of 100 animals at 100 sites, but this time of bonobos. All of the parameters determining bonobo behavior were kept the same as with the chimpanzees, except that we set the bonobo neophobia scores to be, on average, just under one standard deviation higher than the chimpanzee neophobia scores (specifically, this was set as the species difference being 1.5 times larger than the between-site standard deviation, such that:

$$\begin{gathered} Neophobia_{bonobo} \\ \beta_0 = 950 \\ L_{0l} \sim N(0, 100) \\ S_{0s} \sim N(0, 100) \\ e_{ls} \sim N(0, 50) \end{gathered}$$

The decision to make bonobos more neophobic than chimpanzees was arbitrary, and most empirical data supports the opposite conclusion (e.g., Forss et al., 2019). The average self-control scores were kept the same between species. Just as with the replication, we simulated all 10,000 chimpanzees and bonobos, and selected a site at random from which we sampled 10 chimpanzees, and a random site from which we sampled 10 bonobos. Figure 4 shows the results: the entire population of 10,000 chimpanzees in grey circles and 10,000 bonobos in blue circles, and our samples are highlighted.

**Figure 4**

*A Comparison Between Hypothetical Samples of Chimpanzees and Bonobos*



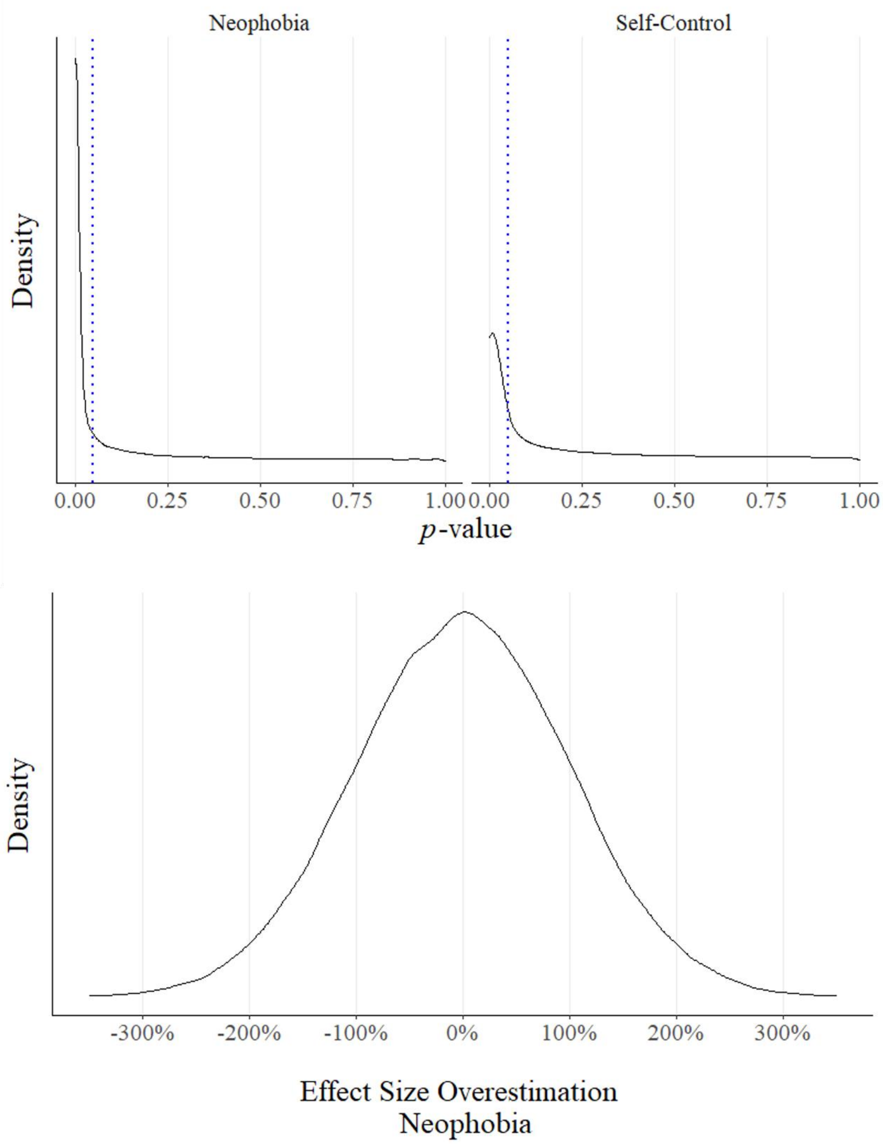
*Note.* Populations of 10,000 chimpanzees (light blue) and 10,000 bonobos (gray) sampled from 100 simulated sites. Samples of 10 chimpanzees and 10 bonobos from a single site are overlaid for chimpanzees (blue) and bonobos (dark grey).

Our samples in Figure 4 captured the direction of the population difference in neophobia scores, which were statistically significantly larger in the bonobo sample than the chimpanzees, $p_{\text{neophobia}} < .001$. However, the magnitude of this effect was overestimated by 41%. For self-control, where no population differences were simulated, our samples produced a statistically significant difference between chimpanzees and bonobos ($p_{\text{self-control}} < .001$), incorrectly estimating a species difference of over 40%. This highlights how even when a statistically significant difference is observed between species at different sites, it does not mean that the difference should be attributed to species identity alone. To explore this further, we investigated how often our comparison would return a statistically significant difference between the neophobia scores and self-control scores of our chimpanzee and bonobo samples. Because our simulation specified that there were no true differences between the species in self-control, this can provide our base-rate of false positive results, under the assumption that statistically significant results would be taken as evidence for a species difference. We simulated 100,000 comparisons between samples of 10 chimpanzees and 10 bonobos, each taken from a new site.

Across the 100,000 simulated comparisons, our small sample design detected a true difference between chimpanzees and bonobos in neophobia 66% of the time with alpha = .05, which looks quite promising. However, the 100,000 simulations also detected a difference between the chimpanzees and bonobos on the self-control measure 49% of the time, in which there were no species differences specified. Figure 5 (upper panel) plots the *p*-value distributions of the two comparisons, and the similarity between these distributions shows that observing a statistically significant difference between two samples, even if $p < .05$, is not necessarily strong evidence of an overall species difference. Figure 5 (lower panel) displays the degree of over- and under-estimation of the neophobia effect size across all samples. Strikingly, in 32% of comparisons, the effect size was overestimated or underestimated by over 100%.

**Figure 5**

*p-value Distributions and Effect Size Overestimation from Two Simulated Comparisons*



*Note.* Upper panel: *p*-value density distributions of two-sample *t*-tests from 100,000 comparisons between 10 hypothetical chimpanzees and 10 hypothetical bonobos, sampled at different sites. The simulation included between-site variation, and a species difference in neophobia, but not self-control. Lower panel: The density distribution of effect size overestimation for the 100,000 comparisons of neophobia behavior. No data are shown for self-control as the set difference was 0, therefore it was not possible to calculate the % overestimation for simulations with non-zero differences.

## Strong and Weak Comparisons

Non-representative sampling leads to weak comparisons, and these comparisons are particularly troublesome when:

- There is a large ratio of within-species variation to between-species variation (MacLean et al., 2012), and absolute species differences are small. Such a scenario will mean the direction and magnitude of differences between samples will be volatile.
- Experimental units are not tested across samples of the same relative settings, measurements, and treatments, and because of this, measurement techniques systematically differ between research programs; For example, when a single population of experimental units is repeatedly sampled, or the same researchers and research groups perform most of the research, with the same experimental designs (Clark et al., 2019; Ioannidis, 2012a; Leavens et al., 2019). This could lead to highly replicable – within narrow boundary conditions - differences between samples being recorded, but these differences being a consequence of specific local features (often confounders) rather than general species differences.

In contrast, strong between-group comparisons should fulfill the following three criteria:

1) The results are consistent within experimental units across times, experimenters, treatments, and measurements within the claims' boundary conditions.
2) The samples of experimental units being compared are tested from within the same relative populations of settings, treatments, and measurements relative to the claim.
3) The between-group differences can be replicated when resampling from the target populations of experimental units.

## Improving Sampling in Animal Research

There are several methods researchers can use to assess and model the effects of biased sampling on the reliability and generalizability of their research findings, which we have divided into experimental design and statistical methods.

### *Experimental Design*

***Increasing Heterogeneity.*** Increasing heterogeneity is a direct method of increasing the representativeness of a sample to a target population. By sampling more diversely from within the populations specified by a theory or claim, researchers can better estimate the population parameters of interest (Milcu et al., 2018; Voelkl et al., 2018, 2020; von Kortzfleisch et al., 2020). This could involve sampling from multiple sites, such as in large collaborative studies (Crabbe et al., 1999; Culina, Adriaensen et al., 2020; Many Primates, Altschul, Beran, Bohn, Call et al., 2019), but also by using multiple different experimenters and varying the conditions and treatments within sites (Baribault et al., 2018; Richter et al., 2010; Würbel, 2002). As an example, Rössler et al. (2020) compared the ability of a sample of wild-caught Goffin's cockatoos and a sample of laboratory-housed Goffin's cockatoos to physically manipulate an apparatus to access a reward. However, rather than presenting the cockatoos with a single apparatus, they were tested in an area with a total of 20 apparatuses. Because Rössler et al. sampled from a diverse range of treatments, we can be confident that - at least for these samples of cockatoos – the results are robust across variations in treatment. An ideal experiment might generate diverse samples across all feasible factors – sites, treatments, experiments, times of day, measurements etc., which will increase the replicability and generalizability of the results (Würbel, 2000); however, it is high-cost (Davies & Gray, 2015; Mook, 1983; Schank & Koehnle, 2009).

***Increasing Homogeneity and Control.*** In contrast to increasing heterogeneity, a lower-cost approach is to increase standardization and control. For example, performing experiments with blinded experimenters only is more homogeneous than performing experiments with a mixture of blinded and non-blinded experimenters. From the re-sampling perspective, blinded and unblinded experimenters come from different populations, and most theories in comparative cognition make predictions that are independent of experimenter bias (i.e., do not predict that experimenter effects are essential for their predictions to be true). Similarly, homogeneity can be useful when a theory is most effectively tested within a subset of the populations that it might apply to. For example, animals are often trained before being tested when researchers attempt to isolate individual psychological mechanisms, such as learning. Such researchers are not usually interested in measuring variability due to neophobia or novel-object exploration, and so animals are familiarized with and trained on the task set-up before being tested to avoid including this "noise" in the dataset. The training pulls all individuals towards their theoretical maximum, increasing statistical power and the relevance of the collected data to the theory in question (Schank & Koehnle, 2009; Smith & Little, 2018), and this can increase the validity of between-group comparisons when the groups have markedly different learning histories (Leavens et al., 2019).

### *Statistical Approaches: Multilevel Models, Phylogenetic Models, and Being Cautious*

The variation in experimental units, settings and measurements can be modeled statistically, using multilevel models (e.g., DeBruine & Barr, 2019; McElreath, 2016), and these should include phylogenetic information for multi-species datasets (Cinar et al., 2020; Davies et al., 2020; Freckleton, 2009; MacLean et al., 2012; Stone et al., 2011). Perhaps most useful are these models' ability to pool information across species and shrink extreme species estimates towards the mean response for a given clade, but it also has the benefit of more closely aligning research fields with evolutionary theory (MacLean et al., 2012; Vonk & Shackelford, 2012). However, generating appropriate multilevel or evolution-informed models of animal behavior is a complex task, which will require a decent amount of data and knowledge about how traits may have been selected. Often, these data and this knowledge will not be available.

When researchers cannot introduce or model variation in their designs, they are faced with a dilemma. Uncertainty intervals will be too narrow with respect to the researcher's populations of interest, but the researcher has no direct means of estimating by how much. One solution is for researchers to artificially increase the uncertainty in their statistical estimates (Kafkafi et al., 2017; Yarkoni, 2019), and this could be informed by data on the ratio of between-site to between-species variance from similar multi-site studies; however, this introduces a trade-off between statistical power and false positive discovery rates. In general, researchers should be cautious when interpreting extreme results observed from single samples, such as the 80% great tit performance on the cylinder task we saw earlier, which regressed to around 40% upon resampling.

### *Barriers*

Concerns about replicability and representativeness have surfaced often in animal behavior and cognition research, at a variety of levels (Beach, 1950; Beran, 2012; Bitterman, 1960; Boesch, 2012, 2020; Brosnan et al., 2013; Clark et al., 2019; Dacey, 2020; Eaton et al., 2018; Farrar et al., 2020; Janmaat, 2019; Leavens et al., 2019; Schubiger et al., 2019; Stevens, 2017; Szabó et al., 2017; van Wilgenburg & Elgar, 2013; Vonk, 2019). However, it is unclear whether any real progress has been made towards understanding the prevalence and consequences of low representativeness in these fields, and we suggest that there are four main reasons why, which are theoretical, practical, motivational, and educational (see also Farrar & Ostojić, 2020).

First, *theoretically*, researchers may believe that their samples are representative of their target populations, or that if they are not, that this does not heavily impact the validity of their results. Such a position may be justifiable, for example when, i) relatively independent animals can be sampled by the same team (e.g., with dog research, or serially captured and released samples), ii) animals are highly trained (Leavens et al., 2019; Skinner, 1956; Smith & Little, 2018), iii) unique case studies, and iv) when heterogenization is used. However, if researchers do justify the generalizability of their findings theoretically, then these arguments should be made explicitly within papers (Simons et al., 2017), be solicited by editors and reviewers (Webster & Rutz, 2020), or provided as commentaries on entire research programmes. These justifications will be strongest when they employ a sampling approach to experimental design, and do not excessively focus on the experimental unit over other levels of sampling variance (Farrar & Ostojić, 2020).

Second, researchers may not *practically* have access to the resources needed to test the representativeness of their samples. They may only have one sample, and other laboratories with access to the same species might not exist. This is a problem – it may not be possible to study hard-to-reach samples in a reliable or replicable manner (Lange, 2019; Leonelli, 2018). However, researchers with such samples can take steps to ensure that their results are as robust as possible, and that an appropriate amount of uncertainty is disclosed, through the experimental and statistical techniques we have mentioned in this article, and so practical constraints do not inherently bar researchers from addressing issues of representativeness.

Third, researchers may lack the *motivation or incentives* to test the representativeness of their samples, and the stability of their results across experimental units, settings, treatments, and measurements. If the scientific incentive and funding structure selects for compelling narratives, oversold findings, and ground-breaking results (Higginson & Munafò, 2016; Ioannidis, 2012b; Smaldino & McElreath, 2016) over rigor, self-correction and understanding, the comparative researcher who attempts to replicate their findings across experimental units and settings may be disadvantaged in terms of common scientific metrics (citations and publications). Addressing these incentive problems is a large task which requires action at the level of the individual (Yarkoni, 2018), organization (Nosek et al., 2012) and society (Amann, 2003; Lazebnik, 2018). Encouragingly, there appears to be a desire to perform more replication studies, in some fields. Fraser et al. (2020), for example, surveyed 439 ecologists, and found that researchers thought replications are very important, reflect a "crucial" use of resources, and should be published by all journals.

Fourth, researchers may be unaware or have not accessed the *education* needed to effectively consider and model sampling variability in their studies. Statistical misconceptions (Goodman, 2008; Hoekstra et al., 2014) and mis-practice (Hoekstra et al., 2012; Nieuwenhuis et al., 2011) are prevalent, present in textbooks (Price et al., 2020), and are perhaps only more likely with the increasing complexity of statistical procedures and software that are available (Forstmeier & Schielzeth, 2011; Schielzeth & Forstmeier, 2009; Silk et al., 2020). At the same time, many university programmes may lack teaching on replication related topics (TARG Meta-Research Group, 2020), and there are no requirements to continue education for researchers following formal qualifications, i.e., post PhD, and neither has considering the replicability or generalizability of findings been well integrated with much of the publishing system (Neuliep & Crandall, 1990, 1993; Webster & Rutz, 2020).

These four barriers will be effectively combatted by top-down measures, such as funding bodies and institutions signing initiatives like the San Francisco Declaration on Research Assessment (DORA), and providing contracts and the job-security needed to promote researchers' scientific development over output metrics. However, bottom-up initiatives from within animal behavior research could be effective and are at least under researchers' direct control (Yarkoni, 2018), and individuals can address each of the four barriers above by, and helping others in, i) discussing how their sampling plans relate to their research aims, and describing what these research aims are, ii) discussing the ethical and practical constraints on diversifying their sampling plans if this is desirable, and considering changes to research designs and generating collaborations if the benefits could outweigh the costs, iii) examining their own motivations when performing science and publishing research findings and, iv) actively pursuing further education in research design and statistical analyses.

## Conclusions

In this article, we applied a resampling definition of replication to animal cognition, and we explored the consequences of small and non-independent (poorly representative) samples in animal behavior and cognition research. Limited sampling is likely a large constraint on the replicability and generalizability of research findings, and it has particularly concerning implications for the accuracy of between group or between species inferences. Comparative researchers should be especially concerned about a lack of representativeness of their samples when there is a large ratio of within-species variation to between-species variation, and when the same researchers, animals, and research methods are used repeatedly. Finally, we discussed how researchers can use techniques such as heterogenization, homogenization, and statistical modeling to improve the replicability and representativeness of their results, and considered the practical, theoretical, and motivational factors that might prevent a full assessment of reliability and representativeness in the field.

## Acknowledgements

We would like to thank Marta Halina, Ljerka Ostojić and Piero Amodio for helpful comments throughout an earlier version the manuscript, and the editor, an anonymous reviewer and Alfredo Sánchez-Tójar for helpful suggestions and critiques throughout the review process. BGF was supported by the University of Cambridge BBSRC Doctoral Training Programme (BB/M011194/1).

## References

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

Almeling, L., Hammerschmidt, K., Sennhenn-Reulen, H., Freund, A. M., & Fischer, J. (2016). Motivational shifts in aging monkeys and the origins of social selectivity. *Current Biology*, *26*, 1744–1749. https://doi.org/10.1016/j.cub.2016.04.066

Amann, R. (2003). A Sovietological view of modern Britain. *The Political Quarterly*, *74*, 468–480. https://doi.org/10.1111/1467-923X.00558

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

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature News*, *533*, 452. https://doi.org/10.1038/533452a

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*, 2607–2612. https://doi.org/10.1073/pnas.1708285114

Bausman, W., & Halina, M. (2018). Not null enough: Pseudo-null hypotheses in community ecology and comparative psychology. *Biology and Philosophy*, *33*, 30. https://doi.org/10.1007/s10539-018-9640-4

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

Beran, M. J. (2012). Did you ever hear the one about the horse that could count? *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00357

Beran, M. J. (2015). The comparative science of "self-control": What are we talking about? *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00051

Bitterman, M. E. (1960). Toward a comparative psychology of learning. *American Psychologist*, *15*, 704–712. https://doi.org/10.1037/h0048359

Bliss-Moreau, E., & Baxter, M. G. (2019). Interest in non-social novel stimuli as a function of age in rhesus monkeys. *Royal Society Open Science*, *6*, 182237. https://doi.org/10.1098/rsos.182237

Boesch, C. (2007). What makes us human (*Homo sapiens*)? The challenge of cognitive cross-species comparison. *Journal of Comparative Psychology*, *121*, 227–240. https://doi.org/10.1037/0735-7036.121.3.227

Boesch, C. (2012). The ecology and evolution of social behavior and cognition in primates. In J. Vonk & T. K. Shackleford (Eds.), *The Oxford handbook of comparative evolutionary psychology*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199738182.013.0026

Boesch, C. (2020). Listening to the appeal from the wild. *Animal Behavior and Cognition*, *7*, 257–263. https://doi.org/10.26451/abc.07.02.15.2020

Bohlen, M., Hayes, E. R., Bohlen, B., Bailoo, J., Crabbe, J. C., & Wahlsten, D. (2014). Experimenter effects on behavioral test scores of eight inbred mouse strains under the influence of ethanol. *Behavioural Brain Research*, *272*, 46–54. https://doi.org/10.1016/j.bbr.2014.06.017

Brosnan, S. F., Beran, M. J., Parrish, A. E., Price, S. A., & Wilson, B. J. (2013). Comparative approaches to studying strategy: Towards an evolutionary account of primate decision making. *Evolutionary Psychology*, *11*, 147470491301100320. https://doi.org/10.1177/147470491301100309

Calisi, R. M., & Bentley, G. E. (2009). Lab and field experiments: Are they the same animal? *Hormones and Behavior*, *56*, 1–10. https://doi.org/10.1016/j.yhbeh.2009.02.010

Chang, W. (2014). *extrafont*: *Tools for using fonts*. (R package version 0.17) [Computer software]. https://CRAN.R-project.org/package=extrafont

Cibulski, L., Wascher, C. A. F., Weiß, B. M., & Kotrschal, K. (2014). Familiarity with the experimenter influences the performance of Common ravens (*Corvus corax*) and Carrion crows (*Corvus corone corone*) in cognitive tasks. *Behavioural Processes*, *103*, 129–137. https://doi.org/10.1016/j.beproc.2013.11.013

Cinar, O., Nakagawa, S., & Viechtbauer, W. (2020). Phylogenetic multilevel meta-analysis: A simulation study on the importance of modeling the phylogeny. *EcoEvoRxiv*. https://doi.org/10.32942/osf.io/su4zv

Clark, H., Elsherif, M. M., & Leavens, D. A. (2019). Ontogeny vs. phylogeny in primate/canid comparisons: A meta-analysis of the object choice task. *Neuroscience & Biobehavioral Reviews*, *105*, 178–189. https://doi.org/10.1016/j.neubiorev.2019.06.001

Coomes, J. R., Davidson, G. L., Reichert, M. S., Kulahci, I. G., Troisi, C. A., & Quinn, J. L. (2020). Inhibitory control, personality, and manipulated ecological conditions influence foraging plasticity in the great tit. *BioRxiv*, 2020.12.16.423008. https://doi.org/10.1101/2020.12.16.423008

Crabbe, J. C., Wahlsten, D., & Dudek, B. C. (1999). Genetics of mouse behavior: Interactions with laboratory environment. *Science*, *284*, 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., Adriaensen, F., Bailey, L. D., Burgess, M. D., Charmantier, A., Cole, E. F., Eeva, T., Matthysen, E., Nater, C. R., Sheldon, B. C., Sæther, B.-E., Vriend, S. J. G., Adamík, P., Aplin, L. M., Angulo, E., Artemyev, A., Barba, E., Barišić, S., Belda, E., …Visser, M. E. (2020). Connected data landscape of long-term ecological studies: The SPI-Birds data hub. *EcoEvoRxiv*. https://doi.org/10.32942/osf.io/6gea7

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*, e3000763. https://doi.org/10.1371/journal.pbio.3000763

Dacey, M. (2020). Anecdotal experiments: Evaluating evidence with few animals. *PhilSci-Archive* http://philsci-archive.pitt.edu/17683/

Davidson, G. L., Cooke, A. C., Johnson, C. N., & Quinn, J. L. (2018). The gut microbiome as a driver of individual variation in cognition and functional behaviour. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*, 20170286. https://doi.org/10.1098/rstb.2017.0286

Davies, A. D., Lewis, Z., & Dougherty, L. R. (2020). A meta-analysis of factors influencing the strength of mate-choice copying in animals. *Behavioral Ecology*, *31*, 1279–1290. https://doi.org/10.1093/beheco/araa064

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*, 5295–5304. https://doi.org/10.1002/ece3.1782

DeBruine, L. M., & Barr, D. J. (2019). Understanding mixed effects models through data simulation *PsyArXiv*. https://doi.org/10.31234/osf.io/xp5cy

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

Farmer, H. L., Murphy, G., & Newbolt, J. (2019). Change in stingray behaviour and social networks in response to the scheduling of husbandry events. *Journal of Zoo and Aquarium Research*, *7*, 203–209. https://doi.org/10.19227/jzar.v7i4.441

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

Farrar, B G., & Ostojić, L. (2019). The illusion of science in comparative cognition. *PsyArXiv*. https://doi.org/10.31234/osf.io/hduyx

Farrar, B. G., & Ostojić, L. (2020). It's not just the animals that are STRANGE. *Learning & Behavior*. https://doi.org/10.3758/s13420-020-00442-5

Farrell, S., & Lewandowsky, S. (2010). Computational models as aids to better reasoning in psychology. *Current Directions in Psychological Science*, *19*, 329–335. https://doi.org/10.1177/0963721410386677

Fawcett, G. L., Dettmer, A. M., Kay, D., Raveendran, M., Higley, J. D., Ryan, N. D., Cameron, J. L., & Rogers, J. (2014). Quantitative genetics of response to novelty and other stimuli by infant rhesus macaques (*Macaca mulatta*) across three behavioral assessments. *International Journal of Primatology*, *35*, 325–339. https://doi.org/10.1007/s10764-014-9750-z

Ferreira, V. H. B., Reiter, L., Germain, K., Calandreau, L., & Guesdon, V. (2020). Uninhibited chickens: Ranging behaviour impacts motor self-regulation in free-range broiler chickens (*Gallus gallus domesticus*). *Biology Letters*, *16*, 20190721. https://doi.org/10.1098/rsbl.2019.0721

Forss, S. I. F., Motes-Rodrigo, A., Hrubesch, C., & Tennie, C. (2019). Differences in novel food response between Pongo and Pan. *American Journal of Primatology*, *81*, e22945. https://doi.org/10.1002/ajp.22945

Forstmeier, W., & Schielzeth, H. (2011). Cryptic multiple hypotheses testing in linear models: Overestimated effect sizes and the winner's curse. *Behavioral Ecology and Sociobiology*, *65*, 47–55. https://doi.org/10.1007/s00265-010-1038-5

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

Freckleton, R. P. (2009). The seven deadly sins of comparative analysis. *Journal of Evolutionary Biology*, *22*, 1367–1375. https://doi.org/10.1111/j.1420-9101.2009.01757.x

Geach, P. T. (1973). Ontological relativity and relative identity. In M. K. Munitz (Ed.), *Logic and ontology* (pp. 287–302). New York University Press.

Gómez, O. S., Juristo, N., & Vegas, S. (2010). Replications types in experimental disciplines. *Proceedings of the 2010 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*, 1–10. https://doi.org/10.1145/1852786.1852790

Goodman, S. (2008). A dirty dozen: Twelve P-value misconceptions. *Seminars in Hematology*, *45*, 135–140. https://doi.org/10.1053/j.seminhematol.2008.04.003

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

Hemmer, B. M., Parrish, A. E., Wise, T. B., Davis, M., Branham, M., Martin, D. E., & Templer, V. L. (2019). Social vs. nonsocial housing differentially affects perseverative behavior in rats (*Ratus norvegicus*). *Animal Behavior and Cognition*, *6*, 168–178. https://doi.org/10.26451/abc.06.03.02.2019

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, *466*, 29–29. https://doi.org/10.1038/466029a

Heyes, C., & Dickinson, A. (1990). The intentionality of animal action. *Mind & Language*, *5*, 87–103. https://doi.org/10.1111/j.1468-0017.1990.tb00154.x

Higginson, A. D., & Munafò, M. R. (2016). Current incentives for scientists lead to underpowered studies with erroneous conclusions. *PLOS Biology*, *14*, e2000995. https://doi.org/10.1371/journal.pbio.2000995

Hill, H. M., Yeater, D., Gallup, S., Guarino, S., Lacy, S., Dees, T., & Kuczaj, S. (2016). Responses to familiar and unfamiliar humans by belugas (*Delphinapterus leucas*), bottlenose dolphins (*Tursiops truncatus*), & Pacific white-sided dolphins (*Lagenorhynchus obliquidens*): A replication and extension. *International Journal of Comparative Psychology*, *29*. https://escholarship.org/uc/item/48j4v1s8

Hoekstra, R., Kiers, H. A. L., & Johnson, A. (2012). Are assumptions of well-known statistical techniques checked, and why (not)? *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00137

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, *21*, 1157–1164. https://doi.org/10.3758/s13423-013-0572-3

Höttges, N., Hjelm, M., Hård, T., & Laska, M. (2019). How does feeding regime affect behaviour and activity in captive African lions (*Panthera leo*)? *Journal of Zoo and Aquarium Research*, *7*, 117–125. https://doi.org/10.19227/jzar.v7i3.392

Hurlbert, S. H. (1984). Pseudoreplication and the design of ecological field experiments. *Ecological Monographs*, *54*, 187–211. https://doi.org/10.2307/1942661

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

Ioannidis, J. P. A. (2012b). Why science is not necessarily self-correcting. *Perspectives on Psychological Science*, *7*, 645–654. https://doi.org/10.1177/1745691612464056

Isaksson, E., Utku Urhan, A., & Brodin, A. (2018). High level of self-control ability in a small passerine bird. *Behavioral Ecology and Sociobiology*, *72*, 118. https://doi.org/10.1007/s00265-018-2529-z

Janmaat, K. R. L. (2019). What animals do not do or fail to find: A novel observational approach for studying cognition in the wild. *Evolutionary Anthropology: Issues, News, and Reviews*, *28*, 303–320. https://doi.org/10.1002/evan.21794

Johnson, Z., Brent, L., Alvarenga, J. C., Comuzzie, A. G., Shelledy, W., Ramirez, S., Cox, L., Mahaney, M. C., Huang, Y.-Y., Mann, J. J., Kaplan, J. R., & Rogers, J. (2015). Genetic influences on response to novel objects and dimensions of personality in Papio baboons. *Behavior Genetics*, *45*, 215–227. https://doi.org/10.1007/s10519-014-9702-6

Kabadayi, C., Bobrowicz, K., & Osvath, M. (2018). The detour paradigm in animal cognition. *Animal Cognition*, *21*, 21–35. https://doi.org/10.1007/s10071-017-1152-0

Kabadayi, C., Krasheninnikova, A., O'Neill, L., van de Weijer, J., Osvath, M., & von Bayern, A. M. P. (2017). Are parrots poor at motor self-regulation or is the cylinder task poor at measuring it? *Animal Cognition*, *20*, 1137–1146. https://doi.org/10.1007/s10071-017-1131-5

Kafkafi, N., Agassi, J., Chesler, E. J., Crabbe, J. C., Crusio, W. E., Eilam, D., Gerlai, R., Golani, I., Gomez-Marin, A., Heller, R., Iraqi, F., Jaljuli, I., Karp, N. A., Morgan, H., Nicholson, G., Pfaff, D. W., Richter, S. H., Stark, P. B., Stiedl, O., …Benjamini, Y. (2018). Reproducibility and replicability of rodent phenotyping in preclinical studies. *Neuroscience & Biobehavioral Reviews*, *87*, 218–232. https://doi.org/10.1016/j.neubiorev.2018.01.003

Kafkafi, N., Golani, I., Jaljuli, I., Morgan, H., Sarig, T., Würbel, H., Yaacoby, S., & Benjamini, Y. (2017). Addressing reproducibility in single-laboratory phenotyping experiments. *Nature Methods*, *14*, 462–464. https://doi.org/10.1038/nmeth.4259

Langbein, J. (2018). Motor self-regulation in goats (*Capra aegagrus hircus*) in a detour-reaching task. *PeerJ*, *6*, e5139. https://doi.org/10.7717/peerj.5139

Lange, F. (2019). Are difficult-to-study populations too difficult to study in a reliable way? *European Psychologist*, *25*(1), 41-50. https://doi.org/10.1027/1016-9040/a000384

Lazebnik, Y. (2018). Who is Dr. Frankenstein? Or, what Professor Hayek and his friends have done to science. *Organisms. Journal of Biological Sciences*, *2*. https://doi.org/10.13133/2532-5876_4_AHEAD1

Lazic, S. E. (2010). The problem of pseudoreplication in neuroscientific studies: Is it affecting your analysis? *BMC Neuroscience*, *11*, 5. https://doi.org/10.1186/1471-2202-11-5

Lazic, S. E. (2016). *Experimental design for laboratory biologists: Maximising information and improving reproducibility* (1st ed.). Cambridge University Press. https://doi.org/10.1017/9781139696647

Leavens, D. A., Bard, K. A., & Hopkins, W. D. (2019). The mismeasure of ape social cognition. *Animal Cognition*, *22*, 487–504. https://doi.org/10.1007/s10071-017-1119-1

Leonelli, S. (2018). Re-thinking reproducibility as a criterion for research quality. *PhilSci-Archive* http://philsci-archive.pitt.edu/14352/

Lewejohann, L., Reinhard, C., Schrewe, A., Brandewiede, J., Haemisch, A., Görtz, N., Schachner, M., & Sachser, N. (2006). Environmental bias? Effects of housing conditions, laboratory environment and experimenter on behavioral tests. *Genes, Brain and Behavior*, *5*, 64–72. https://doi.org/10.1111/j.1601-183X.2005.00140.x

Lewis, D. K. (1993). Many, but almost one. In K. Cambell, J. Bacon, & L. Reinhardt (Eds.), *Ontology, causality and mind: Essays on the philosophy of D. M. Armstrong* (pp. 23–38). Cambridge University Press.

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

Lind, J., Ghirlanda, S., & Enquist, M. (2019). Social learning through associative processes: A computational theory. *Royal Society Open Science*, *6*, 181777. https://doi.org/10.1098/rsos.181777

Lucon-Xiccato, T., Gatto, E., & Bisazza, A. (2017). Fish perform like mammals and birds in inhibitory motor control tasks. *Scientific Reports*, *7*, 1–8. https://doi.org/10.1038/s41598-017-13447-4

Machery, E. (2020). What is a replication? *Philosophy of Science*, *87*. 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*, E2140-2148. https://doi.org/10.1073/pnas.1323533111

MacLean, E. L., Matthews, L. J., Hare, B. A., Nunn, C. L., Anderson, R. C., Aureli, F., Brannon, E. M., Call, J., Drea, C. M., Emery, N. J., Haun, D. B. M., Herrmann, E., Jacobs, L. F., Platt, M. L., Rosati, A. G., Sandel, A. A., Schroepfer, K. K., Seed, A. M., Tan, J., …Wobber, V. (2012). How does cognition evolve? Phylogenetic comparative psychology. *Animal Cognition*, *15*, 223–238. https://doi.org/10.1007/s10071-011-0448-8

MacLean, E. L., Snyder-Mackler, N., vonHoldt, B. M., & Serpell, J. A. (2019). Highly heritable and functionally relevant breed differences in dog behaviour. *Proceedings of the Royal Society B: Biological Sciences*, *286*, 20190716. https://doi.org/10.1098/rspb.2019.0716

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. *Japanese Psychological Review*, *62*, 205–220.

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*, e0223675. https://doi.org/10.1371/journal.pone.0223675

McElreath, R. (2016). *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC Press/Taylor & Francis Group.

Mercado, E. (2016). Commentary: Interpretations without justification: A general argument against Morgan's Canon. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00452

Milcu, A., Puga-Freitas, R., Ellison, A. M., Blouin, M., Scheu, S., Freschet, G. T., Rose, L., Barot, S., Cesarz, S., Eisenhauer, N., Girin, T., Assandri, D., Bonkowski, M., Buchmann, N., Butenschoen, O., Devidal, S., Gleixner, G., Gessler, A., Gigon, A., …Roy, J. (2018). Genotypic variability enhances the reproducibility of an ecological study. *Nature Ecology & Evolution*, *2*, 279–287. https://doi.org/10.1038/s41559-017-0434-x

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

Mook, D. G. (1983). In defense of external invalidity. *American Psychologist*, *38*, 379–387. https://doi.org/10.1037/0003-066X.38.4.379

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

Neuliep, J. W., & Crandall, R. (1993). Reviewer bias against replication research. *Journal of Social Behavior & Personality*, *8*, 21–29.

Nieuwenhuis, S., Forstmann, B. U., & Wagenmakers, E.-J. (2011). Erroneous analyses of interactions in neuroscience: A problem of significance. *Nature Neuroscience*, *14*, 1105–1107. https://doi.org/10.1038/nn.2886

Noonan, H., & Curtis, B. (2004). Identity. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2018 Edition) https://plato.stanford.edu/archives/sum2018/entries/identity/

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

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*, 615–631. https://doi.org/10.1177/1745691612459058

Pfungst, O. (2018). *Clever Hans (the horse of Mr. Von Osten): A contribution to experimental animal and human psychology* (C. L. Rahn, Trans.; 1st ed.) Holt. (Original work published 1911)

Polla, E. J., Grueter, C. C., & Smith, C. L. (2018). Asian elephants (*Elephas maximus*) discriminate between familiar and unfamiliar human visual and olfactory cues. *Animal Behavior and Cognition*, *5*, 279–291. https://doi.org/10.26451/abc.05.03.03.2018

Price, R., Bethune, R., & Massey, L. (2020). Problem with *p* values: Why *p* values do not tell you if your treatment is likely to work. *Postgraduate Medical Journal*, *96*, 1–3. https://doi.org/10.1136/postgradmedj-2019-137079

Quine, W. V. (1950). Identity, ostension, and hypostasis. *The Journal of Philosophy*, *47*, 621–633. https://doi.org/10.2307/2021795

R Core Team (2020). *R: A language and environment for statistical computing*. [Computer software]. R Foundation for Statistical. https://www.R-project.org/

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (2nd ed., pp. 64–99). Appleton Century Crofts.

Richter, S. H., Garner, J. P., Auer, C., Kunert, J., & Würbel, H. (2010). Systematic variation improves reproducibility of animal experiments. *Nature Methods*, *7*, 167–168. https://doi.org/10.1038/nmeth0310-167

Richter, S. H., Garner, J. P., & Würbel, H. (2009). Environmental standardization: Cure or cause of poor reproducibility in animal experiments? *Nature Methods*, *6*, 257–261. https://doi.org/10.1038/nmeth.1312

Richter, S. H., Garner, J. P., Zipser, B., Lewejohann, L., Sachser, N., Touma, C., Schindler, B., Chourbaji, S., Brandwein, C., Gass, P., van Stipdonk, N., van der Harst, J., Spruijt, B., Võikar, V., Wolfer, D. P., & Würbel, H. (2011). Effect of population heterogenization on the reproducibility of mouse behavior: A multi-laboratory study. *PLOS ONE*, *6*, e16461. https://doi.org/10.1371/journal.pone.0016461

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*, 358–367. https://doi.org/10.1037/0033-295X.107.2.358

Rössler, T., Mioduszewska, B., O'Hara, M., Huber, L., Prawiradilaga, D. M., & Auersperg, A. M. I. (2020). Using an Innovation Arena to compare wild-caught and laboratory Goffin's cockatoos. *Scientific Reports*, *10*, 8681. https://doi.org/10.1038/s41598-020-65223-6

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

Schielzeth, H., & Forstmeier, W. (2009). Conclusions beyond support: Overconfident estimates in mixed models. *Behavioral Ecology*, *20*, 416–420. https://doi.org/10.1093/beheco/arn145

Schubiger, M. N., Kissling, A., & Burkart, J. M. (2019). Does opportunistic testing bias cognitive performance in primates? Learning from drop-outs. *PLOS ONE*, *14*, e0213727. https://doi.org/10.1371/journal.pone.0213727

Silk, M. J., Harrison, X. A., & Hodgson, D. J. (2020). Perils and pitfalls of mixed-effects regression models in biology. *PeerJ*, *8*, e9522. https://doi.org/10.7717/peerj.9522

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, *12*, 1123–1128. https://doi.org/10.1177/1745691617708630

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

Skinner, B. F. (1976). *About behaviorism*. Vintage Books.

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*, 160384. https://doi.org/10.1098/rsos.160384

Smith, J. D., Zakrzewski, A. C., & Church, B. A. (2016). Formal models in animal-metacognition research: The problem of interpreting animals' behavior. *Psychonomic Bulletin & Review*, *23*, 1341–1353. https://doi.org/10.3758/s13423-015-0985-2

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

Sorge, R. E., Martin, L. J., Isbester, K. A., Sotocinal, S. G., Rosen, S., Tuttle, A. H., Wieskopf, J. S., Acland, E. L., Dokova, A., Kadoura, B., Leger, P., Mapplebeck, J. C. S., McPhail, M., Delaney, A., Wigerblad, G., Schumann, A. P., Quinn, T., Frasnelli, J., Svensson, C. I., …Mogil, J. S. (2014). Olfactory exposure to males, including men, causes stress and related analgesia in rodents. *Nature Methods*, *11*, 629–632. https://doi.org/10.1038/nmeth.2935

Stevens, J. R. (2017). Replicability and reproducibility in comparative psychology. *Frontiers in Psychology*, 8, 862. https://doi.org/10.3389/fpsyg.2017.00862

Stone, G. N., Nee, S., & Felsenstein, J. (2011). Controlling for non-independence in comparative analysis of patterns across populations within species. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*, 1410–1424. https://doi.org/10.1098/rstb.2010.0311

Stow, M. K., Vernouillet, A., & Kelly, D. M. (2018). Neophobia does not account for motoric self-regulation performance as measured during the detour-reaching cylinder task. *Animal Cognition*, *21*, 565–574. https://doi.org/10.1007/s10071-018-1189-8

Szabó, D., Mills, D. S., Range, F., Virányi, Z., & Miklósi, Á. (2017). Is a local sample internationally representative? Reproducibility of four cognitive tests in family dogs across testing sites and breeds. *Animal Cognition*, *20*, 1019–1033. https://doi.org/10.1007/s10071-017-1133-3

TARG Meta-Research Group. (2020). Statistics education in undergraduate psychology: A survey of UK course content. *PsyArXiv*. https://doi.org/10.31234/osf.io/jv8x3

Tomasello, M., & Call, J. (2008). Assessing the validity of ape-human comparisons: A reply to Boesch (2007). *Journal of Comparative Psychology*, *122*, 449–452. https://doi.org/10.1037/0735-7036.122.4.449

Troisi, C. A., Cooke, A. C., Davidson, G. L., de la Hera, I., Reichert, M. S., & Quinn, J. L. (2020). No evidence for cross-contextual consistency in spatial learning and behavioural flexibility in a passerine. *BioRxiv*, 2020.09.04.282566. https://doi.org/10.1101/2020.09.04.282566

van Wilgenburg, E., & Elgar, M. A. (2013). Confirmation bias in studies of nestmate recognition: A cautionary note for research into the behaviour of animals. *PLOS ONE*, *8*, e53548. https://doi.org/10.1371/journal.pone.0053548

Voelkl, B., Altman, N. S., Forsman, A., Forstmeier, W., Gurevitch, J., Jaric, I., Karp, N. A., Kas, M. J., Schielzeth, H., Van de Casteele, T., & Würbel, H. (2020). Reproducibility of animal research in light of biological variation. *Nature Reviews Neuroscience*, *21*, 1–10. https://doi.org/10.1038/s41583-020-0313-3

Voelkl, B., Vogt, L., Sena, E. S., & Würbel, H. (2018). Reproducibility of preclinical animal research improves with heterogeneity of study samples. *PLOS Biology*, *16*, e2003693. https://doi.org/10.1371/journal.pbio.2003693

von Kortzfleisch, V. T., Karp, N. A., Palme, R., Kaiser, S., Sachser, N., & Richter, S. H. (2020). Improving reproducibility in animal research by splitting the study population into several 'mini-experiments.' *Scientific Reports*, *10*, 16579. https://doi.org/10.1038/s41598-020-73503-4

Vonk, J. (2019). A fish eye view of the mirror test. *Learning & Behavior*, *48*, 193-194. https://doi.org/10.3758/s13420-019-00385-6

Vonk, J., & Shackelford, T. K. (2012). Comparative evolutionary psychology: A united discipline for the study of evolved traits. In J. Vonk, & T. K. Shackleford (Eds.), *The Oxford handbook of comparative evolutionary psychology*. (pp. 547–560). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199738182.013.0029

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*, 283–311. https://doi.org/10.1002/neu.10173

Webster, M. M., & Rutz, C. (2020). How STRANGE are your study animals? *Nature*, *582*, 337–340. https://doi.org/10.1038/d41586-020-01751-5

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, 4, 1686, https://doi.org/10.21105/joss.01686

Wickham H., & Seidel, D. (2020). Scales: Scale functions for visualization. R package version 1.1.1. https://CRAN.R-project.org/package=scales

Würbel, H. (2000). Behaviour and the standardization fallacy. *Nature Genetics*, *26*, 263–263. https://doi.org/10.1038/81541

Würbel, H. (2001). Ideal homes? Housing effects on rodent brain and behaviour. *Trends in Neurosciences*, *24*, 207–211. https://doi.org/10.1016/S0166-2236(00)01718-5

Würbel, H. (2002). Behavioral phenotyping enhanced -- beyond (environmental) standardization. *Genes, Brain and Behavior*, *1*, 3–8. https://doi.org/10.1046/j.1601-1848.2001.00006.x

Yarkoni, T. (2018, October 2). No, it's not the incentives—It's you. [Blog Post]. Retrieved from: https://www.talyarkoni.org/blog/2018/10/02/no-its-not-the-incentives-its-you/

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