ABC 2021, 8(2):97-106
Animal Behavior and Cognition DOI:https://doi.org/10.26451/abc.08.02.01.2021

The Status and Value of Replications in Animal Behavior Science

Katharina F. Brecht[1,*], Edward W. Legg[2], Christian Nawroth[3], Hannah Fraser[4] and Ljerka Ostojić[2]

[1] Institute for Neurobiology, University of Tübingen
[2] Department of Psychology, School of Humanities and Social Sciences, University of Rijeka
[3] Institute of Behavioural Physiology, Leibniz Institute for Farm Animal Biology, Dummerstorf
[4] Interdisciplinary Metaresearch Group, School of BioSciences, University of Melbourne, Victoria
[*] Corresponding author (Email: katharina.brecht@uni-tuebingen.de)

Citation – Brecht, K.F., Legg, E.W., Nawroth, C., Fraser, H. & Ostojić, L. (2021). The status and value of replications in animal behavior science. *Animal Behavior and Cognition, 8*(2), 97-106. https://doi.org/10.26451/abc.08.02.01.2021

Abstract – Replications are widely considered an essential tool to evaluate scientific claims. However, many fields have recently reported that replication rates are low and - when they are conducted - many findings do not successfully replicate. These circumstances have led to widespread debates about the value of replications for research quality, credibility of research findings, and factors contributing to current problems with replicability. This special issue brings together researchers from various areas within the field of animal behavior to offer their perspective on the status and value of replications in animal behavior science.

Keywords – Replication, Inferences, Theory, Methods, Animal Behavior, Animal Cognition

The last decade has seen discussions about the role of replications in evaluating the reliability and credibility of claims in psychological sciences and related fields such as neuroscience (Button et al., 2013; Klapwijk et al., 2021; Mulugeta et al., 2018), economics (Berry et al., 2017; Camerer et al., 2016; Mueller-Langer et al., 2019; Schultze et al., 2019), and ecology (Fidler et al., 2017; Fraser et al., 2020; Kelly, 2019; Nakagawa & Parker, 2015). These discussions were largely inspired by studies reporting that, when put to a test, many findings and well-known effects replicate poorly (Button et al., 2013; Open Science Collaboration, 2015; Munafò, 2009). These efforts have led to large-scale movements to study (e.g., Kelly, 2019, Makel et al., 2012) and combat these issues (e.g., Asendorpf et al., 2013; Chambers, 2013; Nosek & Lakens, 2014). Debates have also moved from an initial focus on statistical and methodological procedures (e.g., Button et al., 2013; Cumming et al., 2004; Scheel et al., 2020) to issues with theory and epistemology in the affected disciplines (e.g., Devezer et al., 2019; Fidler et al., 2018; Fried, 2020). A similar, large-scale development has not yet happened in animal behavior science, although the relevance of replications is considered and discussed by several researchers (e.g., Beran, 2018; Farrar et al., 2020; Stevens, 2017). This special issue presents a starting point for discussions around the status and value of replications in the relatively diverse research fields that study and utilize animal behavior. In doing so, we already have a large body of literature in other disciplines to inform our thinking and discussion. At the same time, research questions relevant to animal behavior may differ from those in other fields and working with animals and measuring their behaviors comes with several unique challenges, such that these discussions may also yield specific perspectives, arguments, and recommendations.

The Status of Replications in Animal Behavior

Despite the apparent significance of replications for any area of scientific inquiry, the replicability of animal behavior research is unclear. Traditionally, comparative research with classic lab animals (rats and pigeons) involved replicating previous work that the current study was inspired by.

However, the growing interest in more diverse species of animals, from fish, to parrots, to great apes, made this convention difficult to follow because these species were less available, took longer to train and test, and in some cases, simply had too few individuals to allow replications on independent samples (Beran, 2018). There are few estimates of the rate of replications in animal behavior research today. The ManyPrimates project (ManyPrimates et al., 2019b) reported that 8.7% (50 out of 574) of primate cognition studies published between 2014 and 2019 tested a different population of the same species with the same methodology, which would generally be considered a direct replication. In the field of ecology and evolution, only 0.023% of open access studies were explicitly described as replications (Kelly, 2019). In a survey, the estimated replication rate for ecology was 10%, and 91% of researchers stated that not enough replications are conducted in ecology (Fraser et al., 2020). Similarly, while 85% of researchers agreed that conducting replications in animal behavior and cognition research is important, the majority of surveyed researchers (70%) also reported that too few studies are explicitly replicating previous work (Farrar et al., 2021a). A range of different factors explaining this (apparent) lack of replications are discussed in the theoretical contributions in this special issue, such as reluctance of funders to finance them, time and resource constraints, and incentive structures (**Farrar et al., 2021b**; **Khan & Wascher, 2021**; **Shaw et al., 2021**).

In addition to the factors mentioned above, which affect different research disciplines to a similar extent, commentators in this special issue discuss the barriers to replication that are more common, although not always exclusive, to animal behavior research, namely small sample sizes, especially for endangered species or species where access is limited due to temporal and physical factors, such as great apes (**Khan & Wascher, 2021**) or certain bird species (**Shaw et al., 2021**), and ethical considerations, such as the 3Rs, reduction, refinement, and replacement (**Nawroth & Gygax, 2021**). Resource constraints are further highlighted by some of the empirical studies. For example, **van Buuren et al. (2021)** replicated the influential Weir et al. (2002) study with Betty, the New Caledonian crow (NCC) that bent and unbent materials to retrieve an out-of-reach reward. In this replication study, 17 NCCs were tested with different tasks in which they were required to either bend material or unbend already bent material to reach the food. To investigate this behavioral flexibility on an individual level, a within-subject design was required. However, in one experiment, the initial within-subject design had to be abandoned due to timing constraints, and the resulting between-subject design reduced the power of the statistical tests. Similarly, **O'Neill et al. (2021)** set out to replicate a study claiming that NCCs can reason about hidden causal agents (Taylor et al., 2012). The original paper's interpretation of this data has been challenged on methodological grounds because crucial controls were missing (Boogert et al., 2013; Dymond et al., 2013). **O'Neill et al. (2021)** not only replicated the procedure of the original study, but also aimed to include those additional control conditions. Consequently, although the authors started off with 14 individuals, the need to assign individuals to different treatments reduced the power within each condition, making it harder to interpret the meaning of the null results. Critically, although limited resources affect original studies, too, replication studies are disproportionately impaired. Due to publication bias, effect sizes of published studies are likely overestimated, and thus, a replication requires more power to be able to replicate (true) effects (Schauer & Hedges, 2021). In addition, due to current incentives, for a single researcher it might make more sense to spend their limited resources on running original experiments than replication studies, which seem to be cited less often than the corresponding original work (Hardwicke et al., 2021; **Khan & Wascher, 2021**).

What is a Replication in Animal Behavior?

Another problem raised by the contributions to this issue, and which might account for the low publication rates of replications, is how to categorize replication studies and interpret their results. The traditional view differentiates between direct and conceptual replications with direct replications acting as tests of reliability and conceptual replications acting as tests of generalizability. The Open Science Collaboration (Open Science Collaboration, 2015, p. aac4716-1) defines a direct replication as "the attempt to recreate the conditions believed sufficient for obtaining a previously observed finding." In contrast, Nosek and Errington (2017, p. 1) define a conceptual replication as a study that uses "a different methodology…to test the same hypothesis." Some of the papers in this special issue fit neatly into this framework. For example, **Lawson et al. (2021**) tested whether yellow warblers' alarm calls varied based on different types of threats to their nests, namely brood parasites and nest predators. Previous studies used visual or combined visual and auditory stimuli, but **Lawson et al. (2021)** used only acoustic playbacks of brood parasites or nest predators to investigate the yellow warblers' response to these nest threats. The use of a different methodology to previous studies would typically mean this experiment would be classified as a conceptual replication and demonstrates that responses for threat stimuli presented in one modality generalize to another.

However, the traditional categorization of direct and conceptual replication is not always suitable. If direct replications are demarcated based on researchers' beliefs about what conditions are sufficient for obtaining the previous finding, then the researchers' judgements rely on a complex theoretical framework underpinning their beliefs and may contain competing theories with similar or varying degrees of empirical support. Such judgements are required for practically every decision made when planning a replication and range from questions about what taxonomic rank constitutes the population in which a behaviour/ability is believed to be present (both **Boyle, 2021**, and **Halina, 2021**, allude to this) to whether placing a speaker 15m from a focal individual is a substantial methodological difference to placing a speaker at a 30m distance (**Salis et al., 2021**). An example of how difficult this decision-making process can be is seen in **Lundgren, Gómez Dunlop et al. (2021)** who ran a study investigating the influence of monoaminergic gene expression on red jungle fowl chicks' interindividual behavioural differences. The authors explicitly noted that there is little consensus about how to measure interindividual differences in animal behavior and that the influence of gene expression can be tested using correlations or through active manipulation (such as knock-out techniques). The combination of the complexity of the systems in question, current understanding, and multiple techniques for investigating the same effect make it difficult to establish which methodological changes could impact the results.

The need to make a range of different decisions regarding methodology is not limited to empirical replications but also impacts attempts to replicate computational models of behaviors. For instance, **Invernizzi and Ruxton (2021)** wrote new code to reproduce the behavior of a computational model of ant behavior. The originally reported model (Franks & Deneubourg, 1997) did not indicate how time was simulated, with one possibility being that they used computational time, a measure that has changed substantially in the ~30 years since the publication of the original paper. Thus, **Invernizzi and Ruxton** chose to simulate time based on the number of rounds that were computed where during a round, each simulated ant moved one unit. Given the absence of information about the original method, it is impossible to judge whether the change should make a substantial difference to the outcome of the model.

However, knowing an original study's method does not necessarily help establish whether a replication can be considered direct or conceptual. For a number of studies, the question of whether a future study's method is believed to be sufficient to replicate the original is dependent on one's theoretical position and current knowledge. For example, **Becker et al. (2021)** tested dogs' perception of the Ebbinghaus-Titchener visual illusion using a spontaneous choice task while previous research employed training-based protocols. In training protocols, the animals learned first to choose the larger (or smaller) of two circles, before being presented with the illusory stimuli during test trials. Thus, a priori, the method used should have had minimal influence on perception per se and on the likelihood of obtaining the previously observed finding. The potential for a replication to be motivated by competing theoretical positions is highlighted by **Silva, Faragó et al. (2021)** who tested whether a Portuguese sample could perceive the emotional state of dog barks and categorize the contexts in which the barks were made. The original study was conducted on a Hungarian sample and, at one level, it was expected that this ability was fundamental and would span a range of human populations. Under this theoretical position, other nationalities should perform similarly to the original Hungarian participants. However, the authors were also motivated by evidence of some cultural differences in emotion recognition that could lead to differences between the performance of the original and new population. Thus, taken at a broad theoretical level, the study may count as a direct replication because such emotion perception is theorized to be universal but, when the different sample is considered in light of differences in emotion recognition between populations, it is more likely to be considered, under traditional views of replication, as either a conceptual replication or an extension of the original research.

The ambiguity created by misaligned theories for replication studies is not simply a definitional issue (e.g., whether a study is a direct or conceptual replication). These issues have important consequences for what might be considered a failed or successful replication. The definition of a "direct replication" states that the conditions used should be believed to be sufficient to obtain the previously observed findings. Critically, theory influences what constitutes obtaining the previous findings in the same way that theory guides the interpretation of what changes to conditions constitute a direct vs conceptual replication. For example, **Lawson et al. (2021)** demonstrate that, as in previous research, yellow warblers are more likely to make seet calls to brood parasites than to nest predators but there was no statistically significant difference in how closely the yellow warblers approached the brood parasites compared to the nest predators, a difference that had been reported in previous research. In this case, the call type was of primary theoretical importance and so the replication can be considered successful. However, it also highlights that, when studies have multiple dependent variables, replications may not always replicate every comparison in the original study. A further issue with interpreting the success of replications is that even the closest possible replication is unlikely to produce a result that precisely matches the results of the previous study but there remain questions over what degree of similarity constitutes success. In light of this question, **Salis et al. (2021)** discuss how the interpretation of whether their results matched the previous study's results largely depends on whether the comparison is based on *p*-values or effect sizes. Relatedly, **O'Neill et al. (2021)** did not find a statistically significant difference between the control and experimental conditions, but it remains unclear whether this is linked to the small sample size/ low power and, consequently, due to type II error. To this end, many of the large-scale replication attempts conducted in psychology and other fields have opted to make comparisons of effect sizes between the original study and replication.

With the ever-increasing prevalence and relevance of replications, the question of what exactly constitutes a (successful) replication will likely remain a topic of discussion. Recently, novel perspectives on the debate have been introduced (e.g., Machery, 2020; Nosek & Errington, 2020). **Farrar et al. (2021b)** and **Halina (2021)** specifically discuss the resampling account (Machery, 2020) as an alternative framework to think about replications in animal behaviour research. The usefulness of this account is that it highlights the similarity between replicability and generalizability of results. Replication studies re-sample from a particular population of participants, but resampling is also possible from populations of all aspects of the original experiment's methodology such as sites, measures, and experimental manipulations. Critically, in all these aspects, researchers are interested in generalizing from their sample to the population. Each of these aspects of an experiment has a specific instance (token) in the original experiment and a replication must use a token from the same class (type or population) of tokens. This means that an experiment is no longer considered a replication if it samples from outside of the original types (populations) used in the original methodology (Machery, 2020). Perhaps the most important aspect, albeit a pragmatic one, of the re-sampling view of replications is that it forces researchers to consider the theoretical basis of their re-sampling because the population that can be re-sampled from must be constrained by statements in the original experiment and/or theory (**Halina, 2021**). For example, a study that is designed to test the hypothesis that pigs can fly is not replicated by a study testing whether fish can fly when the original hypothesis is clearly confined to the population of pigs, the even-toed ungulates from the genus *Sus*.

In view of this, **Halina (2021)** argues that, at least in primate theory of mind research, replications are more common than usually assumed. This is because what are typically considered divergent or novel experiments can be considered as tests of the same underlying hypothesis on the same population. For example, according to Halina's argument, the hypothesis that chimpanzees are sensitive to what conspecifics can perceive has been tested using a range of different measurements, treatments, and populations – that is, sampling from these different methods and populations, these studies (reviewed in Krupenye & Call, 2019) have replicated the original finding that chimpanzees respond to what others orient towards (Hare et al., 2000).

To assess whether Experiment B is a replication of Experiment A, we need to consider the theoretical claim made by Experiment A in order to decide which experimental components are coming from the same population (**Farrar et al., 2021**). For example, in **O'Neill et al.'s (2021)** replication of Taylor et al. (2012), one important control counterbalanced the order of conditions, something that was missing in the original study. This counterbalancing means the replication does not have the same procedure as the original study. However, whether NCCs reason about hidden causal agents should not be influenced by the order of the tests conducted (Boogert et al., 2013; Dymond et al., 2013). Thus, the change in order can be understood as a re-sampling of treatment to test the hypothesis that NCCs reason about the hidden causal agent (the theoretical claim in question sensu **Farrar et al., 2021b**). Another example is the paper by **Salis et al. (2021)** who, unknowingly, replicated a study investigating how great tits perceive order of calls of an allopatric species (Dutour et al., 2020). Both sets of authors collected data in the same territory, but in different breeding seasons. Changes were made in the distance of the playback to the tested birds, and to some of the parameters of the calling sequence presented. Such deviations in methodology are often reasonable. However, it also commonly remains open for debate whether these changes re-sample from the same population (here of treatments) as the original study.

The Value of Replications in Animal Behavior Research

The traditional distinction between direct and conceptual replications often highlights that the type of replication conducted is motivated by different epistemological reasons. Direct replications are thought to test an effect's reliability whereas conceptual replications are considered as testing its generalizability because they involve changes to the conditions such that the explanation of the effect is unlikely to be due to just one specific methodology (Nosek & Errington, 2017). However, one consequence of considering the re-sampling approach is that it is apparent that all replications will inevitably re-sample at least one token, namely, time (this issue is at least implicitly acknowledged in **Farrar et al., 2021b**). Thus, at a minimum, all experiments are a test of an effect's generalizability over time. The resampling approach thus illustrates how the question of replicability is always a question of generalizability (**Farrar et al., 2021b**). The link between replication and generalizability is further highlighted when the terminology from the resampling account is linked with statistical terminology (as explained in **Halina, 2021**) – a token that is resampled is one that is treated as a random factor (Machery, 2020) and thus stands as an example of a broader class (see also Yarkoni, 2020).

However, as highlighted by **Farrar et al. (2021b)** and studies where there are competing theoretical positions, there is always the possibility that the theoretical constraints are interpreted differently by different researchers. There is also a likelihood that theories are updated once a replication's results are known. Thus, **Halina (2021)** argues that replications may further 'theorizing and design.' This can be seen as the reason why Halina's argument diverges from Machery's (2020) view of replications, whereby under the resampling account, one type of replication is not epistemologically superior to another, because it is not important what and how much is being resampled. **Halina (2021)**, on the other hand, argues that replications that resample from only one component are to be distinguished from replications that resample from many components at the same time, because in the latter situation, the core theoretical claim is dependent on a larger number of auxiliary hypotheses. In the case of a failed replication, it becomes necessary to test which hypothesis (including the core claim) led to an incorrect prediction, which is harder to establish in the latter case. Until each of the auxiliary hypotheses has been tested there is nothing to suggest the failed replication actually contradicts the original result.

Critically, the same issue applies to original experiments as well as replications (**Boyle, 2021**) where negative results may be considered the result of some aspect of methodology rather than the absence of an effect or ability. **Boyle (2021)** argues that the issue of auxiliary hypotheses obscuring the interpretation of empirical results is not specific to replications but can be seen throughout the literature. Most notably, negative results are often considered to be the result of the exact methodology used in an experiment rather than the absence of an effect or ability in the sample. This issue is also relevant for positive results, given that data are regularly compatible with multiple theoretical explanations (**Boyle, 2021**). According to **Boyle (2021)**, this theoretical openness and uncertainty means that the role of replications is complicated because both the status of a study as a replication and its acceptance as being successful are theoretically driven - yet there are few commonly accepted theories. In light of this, there may be little benefit in distinguishing replications from original research because they have no different epistemic value.

According to **Boyle (2021)**, this theoretical openness is simply a part of the epistemic circumstances of our field. If we accept that, **Boyle** suggests we could redefine progress in our field in line with Andrews' (2014) suggestion of progress by calibration, where the scientific processes resemble putting together a puzzle. Rather than seeing this as a problem of demarcation and calling into question the 'scientificness' of our field (Chambers, 2013; Schmidt, 2009; Zwaan et al., 2018), **Boyle** suggests that it can constitute our starting point. In doing so, both original studies and replication studies need to be conducted and, importantly, also published, to allow this calibration.

Ways Forward

Unfortunately, as emphasized by **Shaw et al. (2021)** and **Khan and Wascher (2021)**, it is still difficult to receive funding for and to publish replication studies in animal behavior research. Recent developments make us hopeful, however. For example, the last two years have seen more than a half a dozen of papers reporting (successful and unsuccessful) attempts to replicate the influential mirror mark test in corvids (Brecht et al., 2020; Buniyaadi et al., 2020; Clary et al., 2020; Parishar et al., 2021; Soler et al., 2020; Vanhooland et al., 2020; Wang et al., 2020). Moreover, registered replications are becoming more common (for a recent example see Motes-Rodrigo et al., 2021). *Animal Behavior and Cognition* specifically states that empirical replication studies will be published, and authors are encouraged to use the Pre-registered Replication Articles format the journal offers (Beran, 2020). Researchers in the field seem to recognize the need for replication studies (**Farrar et al., 2021b**; Fraser et al., 2020), and multiple publications discuss recommendations for increasing replicability of specific fields (Farrar et al., 2020, for comparative cognition; O'Dea et al., 2021, for ecology and evolutionary biology).

Additional recommendations are also discussed by several authors of this special issue. **Farrar et al. (2021b)** and **Nawroth and Gygax (2021)** argue that conducting studies that increase the heterogeneity of samples will lead to more generalizable and thus replicable results (see also Voelkl et al., 2020; Voelkl & Würbel, 2021; von Kortzfleisch et al., 2020). This has clear benefits for animal behavioral sciences in general but may be of particular importance for fields like animal welfare where policies and treatments need to be applied to a wide range of settings that can vary considerably from lab conditions. It may also be important for fields where no replication studies are possible due to different reasons (for example as discussed in **Shaw et al., 2021**) and where such heterogenization can lead to better inferences (but see **Farrar et al., 2021b**, for a discussion when increasing homogeneity may be a better way forward). Relatedly, there were also calls for the use of multi-lab or multi-setting experiments to help both increase sample sizes for difficult to access species and to improve the heterogeneity of settings (**Khan & Wascher, 2021**). Within the field of primatology, the ManyPrimates project has been founded to facilitate collaborative research, ease replication projects, and test the relationship between ecology, behavior, and cognition (ManyPrimates et al., 2019a, 2019b, 2021). More recently, ManyDogs has been established for similar purposes within canine science and is currently recruiting contributors for the first ManyDogs study on dogs' ability to follow human pointing (https://manydogsproject.github.io/).

Finally, some authors highlight how much of animal cognition research lacks a core theory and many competing perspectives exist (e.g., **Boyle, 2021**), which makes classifying replications and interpreting their results challenging. Researchers may be able to employ multiple approaches in a field with theoretical openness. In some research areas, the theoretical openness may be decreased through theory development, aided by formalization (Allen, 2014; **Farrar et al., 2021b**; Guest & Martin, 2020; Lee et al., 2019; Lind, 2018; Smith et al., 2012; van Rooij & Baggio, 2021; for a study that set out to reproduce a model of ant nest wall building in this issue see **Invernizzi & Ruxton, 2021**). In areas in which this is not possible, claims may have to be adjusted to reflect the epistemic circumstances (Yarkoni, 2020) and criteria other than replications can be employed to evaluate the quality of research and the reliability of claims (Leonelli, 2018).

In conclusion, although barriers remain to conducting and publishing replications it is clear that they are a valuable component of empirical research. This special issue highlights that the value of a replication is more than a single result and that, in practice, replications often have a pragmatic role because they help researchers recognise the theoretical assumptions underpinning their choice of experimental design which pushes forward both experimental design and theory building.

References

Allen, C. (2014). Models, mechanisms, and animal minds. *The Southern Journal of Philosophy*, *52*(S1), 75–97. https://doi.org/10.1111/sjp.12072

Andrews, K. (2014). *The animal mind: An introduction to the philosophy of animal cognition*. Routledge.

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*(2), 108–119. https://doi.org/10.1002/per.1919

Becker, N. Prasar-Shreckengast, S., & Byosiere, S.-E. (2021). Methodological challenges of the spontaneous choice task: Are dogs susceptible to the Ebbinghaus-Titchener illusion? *Animal Behavior and Cognition*, *8*(2), 138-151. https://doi.org/10.26451/abc.08.02.04.2021

Beran, M. J. (2018). Replication and pre-registration in comparative psychology. *International Journal of Comparative Psychology*, *31*. https://doi.org/10.46867/ijcp.2018.31.01.09

Beran, M. J. (2020). Editorial: The value and status of replications in animal behavior and cognition research. *Animal Behavior and Cognition*, *7*, i–iii. https://doi.org/10.26451/abc.07.01.01.2020

Berry, J., Coffman, L. C., Hanley, D., Gihleb, R., & Wilson, A. J. (2017). Assessing the rate of replication in economics. *American Economic Review*, *107*(5), 27–31. https://doi.org/10.1257/aer.p20171119

Boogert, N. J., Arbilly, M., Muth, F., & Seed, A. M. (2013). Do crows reason about causes or agents? The devil is in the controls. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(4), E273. https://doi.org/10.1073/pnas.1219664110

Boyle, A. (2021). Replication, uncertainty and progress in comparative cognition. *Animal Behavior and Cognition*, *8*(2), 295-303. https://doi.org/10.26451/abc.08.02.15.2021

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

Buniyaadi, A., Taufique, S. K. T., & Kumar, V. (2020). Self-recognition in corvids: Evidence from the mirror-mark test in Indian house crows (*Corvus splendens*). *Journal of Ornithology*, *161*(2), 341–350. https://doi.org/10.1007/s10336-019-01730-2

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., & Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433–1436. https://doi.org/10.1126/science.aaf0918

Chambers, C. D. (2013). *Trust in science would be improved by study pre-registration*. https://www.theguardian.com/science/blog/2013/jun/05/trust-in-science-study-pre-registration

Clary, D., Stow, M. K., Vernouillet, A., & Kelly, D. M. (2020). Mirror-mediated responses of California scrub jays (*Aphelocoma californica*) during a caching task and the mark test. *Ethology*, *126(2)*, 140-152. https://doi.org/10.1111/eth.12954

Cumming, G., Williams, J., & Fidler, F. (2004). Replication and researchers' understanding of confidence intervals and standard error bars. *Understanding Statistics*, *3*(4), 299–311. https://doi.org/10.1207/s15328031us0304_5

Devezer, B., Nardin, L. G., Baumgaertner, B., & Buzbas, E. O. (2019). Scientific discovery in a model-centric framework: Reproducibility, innovation, and epistemic diversity. *PLOS ONE*, *14*(5), e0216125. https://doi.org/10.1371/journal.pone.0216125

Dutour, M., Suzuki, T. N., & Wheatcroft, D. (2020). Great tit responses to the calls of an unfamiliar species suggest conserved perception of call ordering. *Behavioral Ecology and Sociobiology*, *74*(3), 37. https://doi.org/10.1007/s00265-020-2820-7

Dymond, S., Haselgrove, M., & McGregor, A. (2013). Clever crows or unbalanced birds? *Proceedings of the National Academy of Sciences*, *110*(5), E336–E336. https://doi.org/10.1073/pnas.1218931110

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., Ostojic, L., & Clayton, N. (2021a). *The hidden side of animal cognition research: Scientists' attitudes toward bias, replicability and scientific practice*. PsyArXiv. https://doi.org/10.31234/osf.io/fwvpe

Farrar, B. G., Voudouris, K., & Clayton, N. S. (2021b). Replications, comparisons, sampling and the problem of representativeness in animal behavior and cognition research. *Animal Behavior and Cognition* *8*(2), 272-294. https://doi.org/10.26451/abc.08.02.14.2021

Fidler, F., Chee, Y. E., Wintle, B. C., Burgman, M. A., McCarthy, M. A., & Gordon, A. (2017). Metaresearch for evaluating reproducibility in ecology and evolution. *Bioscience*, *67*(3), 282–289. https://doi.org/10.1093/biosci/biw159

Fidler, F., Singleton Thorn, F., Barnett, A., Kambouris, S., & Kruger, A. (2018). The epistemic importance of establishing the absence of an effect. *Advances in Methods and Practices in Psychological Science*, *1*(2), 237–244. https://doi.org/10.1177/2515245918770407

Franks, N. R., & Deneubourg, J.-L. (1997). Self-organizing nest construction in ants: Individual worker behaviour and the nest's dynamics. *Animal Behaviour*, *54*(4), 779–796. https://doi.org/10.1006/anbe.1996.0496

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

Fried, E. I. (2020). Lack of theory building and testing impedes progress in the factor and network literature. *Psychological Inquiry*, *31*(4), 271–288. https://doi.org/10.1080/1047840X.2020.1853461

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

Halina, M. (2021). Replications in comparative psychology. *Animal Behavior and Cognition 8*(2), 262-271. https://doi.org/10.26451/abc.08.02.13.2021

Hardwicke, T. E., Szűcs, D., Thibault, R. T., Crüwell, S., Akker, O. van den, Nuijten, M. B., & Ioannidis, J. P. A. (2021). *Post-replication citation patterns in psychology: Four case studies*. MetaArXiv. https://doi.org/10.31222/osf.io/wt5ny

Hare, B., Call, J., Agnetta, B., & Tomasello, M. (2000). Chimpanzees know what conspecifics do and do not see. *Animal Behaviour*. https://doi.org/10.1006/ANBE.1999.1377

Invernizzi, E., & Ruxton, G. D. (2021). Updating a textbook model of collective behaviour: Nest wall building in *Temnothorax albipennis*. *Animal Behavior and Cognition, 8*(2), 230-238. https://doi.org/10.26451/abc.08.02.09.2021

Kelly, C. D. (2019). Rate and success of study replication in ecology and evolution. *PeerJ*, *7*, e7654. https://doi.org/10.7717/peerj.7654

Khan, N., & Wascher, C. (2021). Considering generalizability: A lesson from auditory enrichment research on zoo animals. *Animal Behavior and Cognition, 8*(2), 250-261. https://doi.org/10.26451/abc.08.02.12.2021

Klapwijk, E. T., van den Bos, W., Tamnes, C. K., Raschle, N. M., & Mills, K. L. (2021). Opportunities for increased reproducibility and replicability of developmental neuroimaging. *Developmental Cognitive Neuroscience*, *47*, 100902. https://doi.org/10.1016/j.dcn.2020.100902

Krupenye, C., & Call, J. (2019). Theory of mind in animals: Current and future directions. *WIREs: Cognitive Science*, *10*(6), e1503. https://doi.org/10.1002/wcs.1503

Lawson, S. L., Enos, J. K., Mendes, N. C. Gill, S. A., & Hauber, M. E. (2021). Responses of female Yellow Warblers to playbacks signaling brood parasitism and predation risk: a quasi-replication study. *Animal Behavior and Cognition*, . 8(2), 216-229. https://doi.org/10.26451/abc.08.02.08.2021

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), 141–153. https://doi.org/10.1007/s42113-019-00029-y

Leonelli, S. (2018). Rethinking reproducibility as a criterion for research quality. In (Ed) *Including a Symposium on Mary Morgan: Curiosity, Imagination, and Surprise* (Vol. 36, p. 129–146). Emerald Publishing. https://doi.org/10.1108/S0743-41542018000036B009

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

Lundgren, K. A., Gómez Dunlop, C. A., Garnham, L. C., Ryding, S., Abbey-Lee, R. N., Kreshchenko, A., & Løvlie, H. (2021). Exploring the relationship between boldness, activity, exploration and monoaminergic gene expression in red junglefowl chicks. *Animal Behavior and Cognition, 8*(2), 124-137. https://doi.org/10.26451/abc.08.02.03.2021

Machery, E. (2020). What is a replication? *Philosophy of Science*, *87*(4), 545–567. https://doi.org/10.1086/709701

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science*, *7*(6), 537–542. https://doi.org/10.1177/1745691612460688

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

ManyPrimates, Altschul, D., Bohn, M., Canteloup, C., Ebel, S., Hernandez-Aguilar, R. A., Joly, M., Keupp, S., Petkov, C., Llorente, M., Proctor, D., Motes-Rodrigo, A., Sutherland, K., Szabelska, A., Taylor, D., Völter, C., & Wiggenhauser, N. G. (2021). *Collaboration and Open Science Initiatives in Primate Research*. OSF Preprints. https://doi.org/10.31219/osf.io/7c93a

ManyPrimates, 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. (2019b). Establishing an infrastructure for collaboration in primate cognition research. *PLOS ONE*, *14*(10), e0223675. https://doi.org/10.1371/journal.pone.0223675

Motes-Rodrigo, A., Mundry, R., Call, J., & Tennie, C. (2021). Evaluating the influence of action- and subject-specific factors on chimpanzee action copying. *Royal Society Open Science*, *8*(2), 200228. https://doi.org/10.1098/rsos.200228

Mueller-Langer, F., Fecher, B., Harhoff, D., & Wagner, G. G. (2019). Replication studies in economics—How many and which papers are chosen for replication, and why? *Research Policy*, *48*(1), 62–83. https://doi.org/10.1016/j.respol.2018.07.019

Mulugeta, L., Drach, A., Erdemir, A., Hunt, C. A., Horner, M., Ku, J. P., Myers Jr., J. G., Vadigepalli, R., & Lytton, W. W. (2018). Credibility, replicability, and reproducibility in simulation for biomedicine and clinical applications in neuroscience. *Frontiers in Neuroinformatics*, *12*, 18. https://doi.org/10.3389/fninf.2018.00018

Munafò, M. R. (2009). Reliability and replicability of genetic association studies. *Addiction*, *104*(9), 1439–1440. https://doi.org/10.1111/j.1360-0443.2009.02662.x

Nakagawa, S., & Parker, T. H. (2015). Replicating research in ecology and evolution: Feasibility, incentives, and the cost-benefit conundrum. *BMC Biology*, *13*(1), 88. https://doi.org/10.1186/s12915-015-0196-3

Nawroth, C., & Gygax L. (2021). The legislative, ethical, and conceptual importance of replicability in farm animal welfare science. *Animal Behavior and Cognition 8*(2), 246-249. https://doi.org/10.26451/abc.08.02.11.2021.

Nosek, B. A., & Errington, T. M. (2017). Making sense of replications. *eLife*, *6*, e23383. https://doi.org/10.7554/eLife.23383

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

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology*, *45*, 137. https://doi.org/10.1027/1864-9335/a000192

O'Dea, R. E., Parker, T. H., Chee, Y. E., Culina, A., Drobniak, S. M., Duncan, D. H., Fidler, F., Gould, E., Ihle, M., Kelly, C. D., Lagisz, M., Roche, D. G., Sánchez-Tójar, A., Wilkinson, D. P., Wintle, B. C., & Nakagawa, S. (2021). Towards open, reliable, and transparent ecology and evolutionary biology. *BMC Biology*, *19*(1), 68. https://doi.org/10.1186/s12915-021-01006-3

O'Neill, L., Linder, G. van Buuren, M., & von Bayern. A. M.P. (2021). New Caledonian crows and hidden agents revisited. *Animal Behavior and Cognition*, *8*(2), 166-189. https://doi.org/10.26451/abc.08.02.06.2021

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). https://doi.org/10.1126/science.aac4716

Parishar, P., Mohapatra, A., & Iyengar, S. (2021). Investigating behavioral responses to mirrors and the mark test in adult male zebra finches and house crows. *Frontiers in Psychology*, *12*, 637850. https://doi.org/10.3389/fpsyg.2021.637850

Salis, A., Lena, J.-P., & Lengagne, T. (2021). How subtle protocol choices can affect biological conclusions: Great tits' response to allopatric mobbing calls. *Animal Behavior and Cognition*, *8*(2), 152-165. https://doi.org/10.26451/abc.08.02.05.2021

Schauer, J. M., & Hedges, L. V. (2021). Reconsidering statistical methods for assessing replication. *Psychological Methods*, *26*(1), 127–139. https://doi.org/10.1037/met0000302

Scheel, A. M., Tiokhin, L., Isager, P. M., & Lakens, D. (2020). Why hypothesis testers should spend less time testing hypotheses. *Perspectives on Psychological Science*, 1745691620966795. https://doi.org/10.1177/1745691620966795

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, *13*(2), 90–100. https://doi.org/10.1037/a0015108

Schultze, T., Huber, J., Kirchler, M., & Mojzisch, A. (2019). Replications in economic psychology and behavioral economics. *Journal of Economic Psychology*, *75*, 102199. https://doi.org/10.1016/j.joep.2019.102199

Shaw, R. C., Greggor, A. L., Plotnik, J. M. (2021). The challenges of replicating research on endangered species. *Animal Behavior and Cognition*, *8*(*2*), 239-245. https://doi.org/10.26451/abc.08.02.10.2021

Silva, K., Farago, T., Pongracz, P., Romeiro, P., Lima, M., de Sousa, L. (2021). Humans' ability to assess emotion in dog barks only slightly affected by their country of residence, a replication of Pongrácz et al. (2005) in a Portuguese sample. *Animal Behavior and Cognition,* *8*(2), 107-123. https://doi.org/10.26451/abc.08.02.02.2021

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

Soler, M., Colmenero, J. M., Pérez-Contreras, T., & Peralta-Sánchez, J. M. (2020). Replication of the mirror mark test experiment in the magpie (*Pica pica*) does not provide evidence of self-recognition. *Journal of Comparative Psychology*, *134*(4), 363–371. https://doi.org/10.1037/com0000223

Stevens, J. R. (2017). Replicability and reproducibility in comparative psychology. *Frontiers in Psychology*, *8*, 1–6. https://doi.org/10.3389/fpsyg.2017.00862

Taylor, A. H., Miller, R., & Gray, R. D. (2012). New Caledonian crows reason about hidden causal agents. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(40), 16389–16391. https://doi.org/10.1073/pnas.1208724109

van Buuren, M., Alfredsson, J., Mioduszewska, B., Tebbich, S., & von Bayern, A. M. P. (2021). Bending and unbending of tools by New Caledonian crows (*Corvus moneduloides*) in a problem-solving context. *Animal Behavior and Cognition 8*(2), 190-215. https://doi.org/10.26451/abc.08.02.07.2021

van Rooij, I., & Baggio, G. (2021). Theory before the test: How to build high-verisimilitude explanatory theories in psychological science. *Perspectives on Psychological Science*, 1745691620970604. https://doi.org/10.1177/1745691620970604

Vanhooland, L.-C., Bugnyar, T., & Massen, J. J. M. (2020). Crows (*Corvus corone* ssp.) check contingency in a mirror yet fail the mirror-mark test. *Journal of Comparative Psychology*, *134*(2), 158–169. https://doi.org/10.1037/com0000195

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*(7), 384–393. https://doi.org/10.1038/s41583-020-0313-3

Voelkl, B., & Würbel, H. (2021). A reaction norm perspective on reproducibility. *Theory in Biosciences*. https://doi.org/10.1007/s12064-021-00340-y

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*(1), 16579. https://doi.org/10.1038/s41598-020-73503-4

Wang, L., Luo, Y., Wang, H., Zou, Y., Yao, H., Ullah, S., & Li, Z. (2020). Azure-winged magpies fail to understand the principle of mirror imaging. *Behavioural Processes*, *177*, 104155. https://doi.org/10.1016/j.beproc.2020.104155

Weir, A. A. S., Chappell, J., & Kacelnik, A. (2002). Shaping of hooks in New Caledonian crows. *Science*, *297*(5583), 981. https://doi.org/10.1126/science.1073433

Yarkoni, T. (2020). The generalizability crisis. *Behavioral and Brain Sciences*, 1–37. https://doi.org/10.1017/S0140525X20001685

Zwaan, R., Etz, A., Lucas, R., & Donnellan, B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, 1–50. https://doi.org/10.1017/S0140525X17001972