*Note: This is the accepted version of a paper to be published in Animal Behaviour and Cognition. Please cite the final, published version.*

# Replication, Uncertainty and Progress in Comparative Cognition

Alexandria Boyle[1]

[1]Leverhulme Centre for the Future of Intelligence, University of Cambridge
Center for Science and Thought, University of Bonn

Address for correspondence: Leverhulme Centre for the Future of Intelligence, 16 Mill Lane, Cambridge, CB2 1SB; Email: asb69@cam.ac.uk

*Note: This is the accepted version of a paper to be published in Animal Behaviour and Cognition. Please cite the final, published version.*

**Replication, Uncertainty and Progress in Comparative Cognition**

Abstract: Replications are often taken to play both epistemic and demarcating roles in science: they provide evidence about the reliability of fields' methods and, by extension, about which fields 'count' as scientific. I argue that in a field characterised by a high degree of theoretical openness and uncertainty, like comparative cognition, replications do not sit well in these roles. Like other experiments conducted under conditions of uncertainty, replications are often equivocal and open to interpretation. As a result, they are poorly placed to deliver clear judgments about the reliability of comparative cognition's methods or its scientific bona fides. I suggest that this should encourage us to take a broader view of both the nature of scientific progress and the role of replication in comparative cognition.

Keywords: Replication; Underdetermination; Uncertainty; Experiment; Scientific Progress

## 1. Introduction

Benjamin Farrar and Ljerka Ostojić (2019) argue that comparative cognition suffers from a bias in favour of confirming hypotheses that attribute complex cognitive capacities to nonhuman animals, lacks strategies for disconfirming such hypotheses, and likely produces a high number of false positive results. In short, comparative cognition 'displays all the hallmarks of a science that could […] soon be thrust into a replication crisis' (2019, p. 13). At the same time, they argue, the prevalence of methodological criticism in the field creates an 'illusion' of scientific rigour, obscuring these problems. This adds up to a worry about comparative cognition's scientific *bona fides*: 'comparative cognition uses sciency-sounding methods, but whether these methods work is a scientific question in itself', to which we should devote more attention (2019, p. 14).

How ought one to address the question of whether comparative cognition's scientific methods work – and of whether they are not merely sciency-sounding, but scientific? A natural thought is that one ought to approach it by conducting more replications.[1] Two considerations make this thought appealing. First, it is widely accepted that one function of replication is to occupy precisely this *epistemic role*: that is, evaluating the effectiveness of a field's scientific methods by detecting false positive results. Second, and relatedly, replication is also widely viewed as occupying a *demarcating role*: that is to say, whether a field or its methods count as properly scientific is, at least in part, determined by their replicability. So, if one's interest is in evaluating the effectiveness and scientific credentials of comparative cognition's methods, replication is a natural place to begin.

In what follows, I argue that replication fits poorly in these roles in comparative cognition.[2] As a result of the theoretical openness and uncertainty characterising research in comparative cognition, experiment in comparative cognition is routinely multiply interpretable. As a special case of experiments, the same is true of replications, with the result that it will often be unclear whether a study qualifies as a replication at all. As a result, replications will often give rise to the same kind of methodological criticism associated with experiment generally, yielding no straightforward judgments about the reliability of comparative cognition's methods. This is not to say that replications have no role to play in making progress in comparative cognition. On the view that I suggest, the role of replications is the same as that of experiment in general: to calibrate the range of available theories, rather than to conclusively tell between them.

## 2. Functions for Replication

Theoretical discussions of replication offer a variety of functions for replication, which we might group broadly into epistemic and demarcating functions (e.g. Schmidt, 2009; Zwaan et al., 2018).

First, replications are taken to occupy a variety of epistemic functions: depending on the type of replication, they provide information about the reliability, reality or generalisability of previous studies or effects. It is standard to distinguish two forms of replication, 'direct' and 'conceptual' (but see Machery, 2020). As this distinction is

---

[1] I use the term 'replication' to denote any replication attempt, and use 'successful replication' and 'failed replication' respectively to indicate replication attempts which do and do not yield similar results to those of the original study.

[2] Similar points may apply to other scientific fields, but here I restrict my attention to comparative cognition.

typically construed, direct replications aim to reproduce an experiment as closely as possible, holding all relevant factors fixed, with a view to determining whether the original experiment was reliable – that its result was not the result of non-directional error, questionable research practices and so on (see, e.g. Schmidt, 2009, p. 93). Conceptual replications, on the other hand, aim to test the original hypothesis by other means, with a view to learning more about the nature of the effect uncovered. We might distinguish these in turn from extensions, which explore the validity of experiments and invariance range of their effects.

Second, replicability is also treated as a 'demarcation criterion' for science – that is, a (partial) answer to the question of what makes a field of inquiry scientific. This view of replication takes its inspiration from the work of Karl Popper (Derksen, 2019). As Popper put it, science is concerned with effects which, 'on account of their regularity and reproducibility are inter-subjectively testable' (Popper, 2002, p. 23). Thus, Zwaan et al. (2018, p. 2) write that replicability is 'an essential criterion' for effects being 'accepted as part of the scientific literature'. Similarly, Schmidt (2009, p. 92) writes that replicability functions as a scientific norm: 'any assertion that cannot be demonstrated in a replication is not regarded as a scientific statement'. It should be clear that this demarcating function for replication is related to its epistemic functions: replicability can plausibly serve as a demarcation criterion only because – and to the extent that – it provides information about the reliability of methods, the reality of effects, and so on.

Also in virtue of occupying these epistemic roles, replication is often taken to be intimately related to a second demarcation criterion for science: falsifiability. Popper, again, famously argued that science and scientific progress could be understood in terms of falsification. A scientific theory is one making 'risky' predictions which can be falsified; scientific progress consists not in confirming theories but in falsifying them (2002, sec. 6). Experiments are a tool in making this progress and so, as a special case of experiments, are replications: falsification can be achieved 'via a meticulously executed series of *direct replications'* (LeBel, 2017, cited in Derksen, 2019, p. 452). If an effect from a previous experiment (repeatedly) fails to replicate, this can be taken to falsify the existence of the effect, and the theories that postulate it. Conversely, if an effect does (repeatedly) replicate, we might take this to falsify the claim that it was an isolated coincidence, and any theories that rule out the effect's existence – though a Popperian would stop short of saying that our theory about the effect is confirmed.

That this familiar story about both the epistemic role of replication and the demarcation of science is too simple is illustrated quite vividly by an experiment conducted by CERN in 2011, in which neutrinos appeared to travel from Geneva to Gran Sasso at faster-than-light speeds. This result, which was incompatible with the theory of special relativity, was significant at the six-sigma level, and was directly replicated. But rather than take the result and its replication to falsify special relativity, 'everyone instead set out to discern what was wrong with the experiment' (Lewens, 2015, p. 26). This was a perfectly rational, scientific thing to do. Special relativity is an extraordinarily well-established theory, and that an experiment can be successfully replicated does not make it a good one. Under the circumstances, and even despite the successful replications, it seemed far more likely that the effect was the result of an error – as it indeed was – than that special relativity was wrong.

This example is instructive not because I take particle physics and comparative cognition to have a great deal in common, but because it is helpful to begin with a case whose scientific credentials are not in doubt. It seems uncontroversial that particle physics is a science in good standing, and that the community's response to these results was scientifically respectable. If an account of the epistemic structure and demarcation of science fails to capture this as an instance of science done well, that should reduce our confidence in the adequacy of the account. But we need not stray so far from comparative cognition to find cases that put pressure on this picture. Consider Maes and colleagues' failure to replicate the blocking effect, a 'touchstone' of learning theory (2016, p. e50). Although Maes and colleagues failed to replicate the effect fifteen times, they did not take this string of failed replications to falsify the existence of blocking. On the contrary, they write, 'we have no doubt that true blocking exists' (2016, p. e60). Against a background of well-established theory, another explanation for these results seemed called for – whether in terms of unknown boundary conditions on blocking or problems with the replications' design (see Maes et al., 2018; Soto, 2018). I take no view on which of these explanations is correct; the point is simply that looking for explanations for these replication failures, besides the non-existence of blocking, was a scientifically respectable move.[3]

The point is that experiments and replications do not straightforwardly carry information about the reliability of experiments, the reality of effects or the truth of

[3] Thanks to a reviewer for highlighting the similarity between these cases.

theories. As these cases make clear, their epistemic role in science is more complicated than this. In practice, when experiments and replications disagree with one another or with theories, 'scientists use inductive inference to help them decide where a mistake has most likely been made' (Lewens, 2015, p. 27). But these two examples differ from typical instances of research in comparative cognition in an important respect: in both cases, experiments and replications were conducted against a background of well-established theory which imposed some significant constraints on the interpretation of the evidence. Research in comparative cognition often lacks such a firm theoretical background, instead being characterised by a high degree of uncertainty and 'conceptual openness' (Feest, 2019). As I argue in the next section, this further complicates the epistemic role of replications in comparative cognition.

## 3. Openness and uncertainty

As Farrar and Ostojić (2019) highlight, comparative cognition is a field characterised by widespread methodological and theoretical criticism. When experimental results appear to support the hypothesis that animals have a certain cognitive capacity, critics often offer leaner interpretations of the results (see, for instance, Heyes' (1994) proposal that the results of Gallup's (Gallup, 1970) mark test may have been an artefact of the subjects' anaesthesia) Conversely, when experimental results fail to support a hypothesis, researchers often note a variety of alternative explanations for this besides animals lacking the cognitive capacity of interest (see, for instance, Plotnik and colleagues' (2006) suggestion that some elephants may fail the mark test due to disinterest in their physical appearance, rather than an absence of self-recognition). I suggest that the prevalence of post hoc criticism is understandable given the backdrop of openness and uncertainty against which research in comparative cognition is conducted. Since replications in comparative cognition are conducted in these same epistemic conditions, their results are also equivocal. As a result, replications will often be ill-equipped to deliver clear verdicts about the reliability of comparative cognition's methods or the reality of its effects.

Let us begin by considering experiment in general. Suppose we are interested in investigating a hypothesis like 'population X has capacity C', where X is a population of nonhuman animals, and C is a cognitive capacity we take humans to have. Whilst hypotheses like this appear relatively well-defined, they conceal a multitude of further questions which in many cases lack settled answers.

First, there might be open questions about how to conceptualise C, even in the human case. To take episodic memory, for instance, we might wonder whether it is better conceptualised as an instance of the capacity for mental time travel, or whether it should more centrally be understood in terms of the role it plays in the declarative memory system. Second, we might have questions about C's origins: is it an adaptation, perhaps a developmentally rigid, domain-specific module? Or do learning and environment play an important role in its development – is it perhaps a cognitive gadget (Heyes, 2018)? Third, we might wonder about whether, if C did exist in population X, it would be very similar to the human capacity, or whether it might manifest in species-specific ways. Might selection have favoured some aspects of C over others in X's lineage? Or might C follow a unique developmental trajectory in X? Fourth, for any way of settling these questions about C, there will likely be many ways in which we could operationalise and measure C. Which of these operationalisations and measurements best capture C? Finally, there might be questions about the individuation of the population, X. More often than not, 'X' is the name of a species or higher-level taxonomic unit – something like 'chimpanzees' or 'great apes'. But when we investigate (for instance) whether chimpanzees have C, it is frequently an open question just how general this hypothesis is supposed to be. Are we looking for evidence that C is a stable trait across the chimpanzee species as a whole, or would it be significant if we found evidence of C in some few chimpanzees, perhaps ones with atypical life histories?

The answers to all of these questions are of course related in complicated ways. For instance, if learning and environment play a significant role in the development of a capacity, then we might expect to find evidence of the capacity in some individuals and not others – and conversely, finding evidence of the capacity in just a few individuals might provide evidence about its developmental trajectory (Trestman, 2015). If the capacity is grounded in a domain-specific, developmentally rigid module, then we might instead expect to see it manifest more consistently across a species. How we expect the capacity to manifest in nonhumans, and how we ought to operationalise and measure it, might also turn on what we think about its developmental and evolutionary origins, the likelihood of interspecific variation, and how best to conceptualise the role it plays in the wider cognitive system.

It is important to emphasise that the openness of these questions in no way calls into question comparative cognition's status as a science: this is simply a description of the epistemic circumstances in which science in this area is conducted. There are many open questions about the nature and possible manifestations of cognitive capacities, and about how to individuate the populations that might have them. These open questions are entangled with one another, and with experiment. We necessarily make assumptions about the nature of cognitive capacities when we investigate which animals have them, but which animals have them is also important evidence about the nature of the capacities. Whenever we test the hypothesis that population X has capacity C, we are consequently not only probing this hypothesis, but a multitude of – often unstated – background assumptions about the capacity itself.

As a result, experiment in comparative cognition routinely underdetermines theory. That is to say, experimental results are routinely compatible with a range of theories, and so are insufficient to determine which theoretical conclusions we should draw. Underdetermination is a feature of science generally, since individual hypotheses cannot be tested in isolation but only together with a range of background assumptions or 'auxiliary hypotheses' (Duhem, 1954; Quine, 1958). So, whenever an experiment produces negative results, the fault may be with the hypothesis of interest or with one of these auxiliary ones. By the same token, when an experiment produces positive results, there may be another collection of hypotheses and background assumptions with which it is compatible. The practical consequences of this vary depending on how confident we have reason to be in our background assumptions: the greater our confidence in those assumptions, the more constrained our interpretation of the evidence. In comparative cognition, I am suggesting, there is a high degree of openness and uncertainty surrounding those assumptions. As a result, there are often relatively few constraints on the interpretation of results – making the prevalence of post-experimental criticism in the field unsurprising.

Since replications are a kind of experiment, similar points apply to them: they frequently underdetermine which theoretical conclusions we should draw. But replications give rise to an additional complication, because the conclusions we should draw from a replication study turn on whether and in what sense it is a replication. But openness and uncertainty in the background of an original study often leave it unclear whether and in what sense a later experiment 'counts' as a replication of the original.

Against this background, replication studies are unlikely to yield clear evidence about the reliability of comparative cognition's methods.

To see this, let us imagine that an experiment performed by one research group yields results suggesting that a species has a certain capacity. Subsequently, another group aims to directly replicate this experiment – using the same interventions and the same measurement procedures on other members of the same species. What might we infer from the success or failure of this direct replication attempt?

If the replication does not yield similar results, it is a familiar point that this evidence underdetermines what conclusion we should draw, even in ideal circumstances. When a direct replication attempt disagrees with the original experiment, even assuming all relevant factors were held fixed, this tells us at most that one of the results is a false one: either there is an effect, and the replication was a false negative, or there is no effect, and the original study was a false positive.

But given the openness characterising research in comparative cognition, it is not clear that we are entitled even to this disjunctive conclusion. We could draw that conclusion only if we were sure that, if the effect in the first study was genuine – that is, if the subjects tested did have the capacity – we would find the same effect in the second study. That might be true if the capacity were grounded in a developmentally rigid domain-specific module. But if this is not true of the capacity in question, then it could be that the original effect was genuine, but that the life history of the subjects used by the two groups differs in some important way, such that members of the first group genuinely have the capacity, and members of the second genuinely do not. In this case, there would be a sense in which, although these subjects belong to the same species, they form distinct populations. As such, our second experiment would not qualify as a direct replication at all, since not all relevant factors were held fixed. Instead, we might view it as an extension whose results do not establish anything about reliability, but provide evidence about the nature of the effect and the individuation of the population in which it manifests.

So, the evidence of failed replication is equivocal – and not only in the sense that, when direct replications fail, it is an open question whether the original or the replication is at fault. More importantly than this, the interpretation of replication failure in comparative cognition turns on a host of background assumptions about the nature of cognitive capacities and about what is and is not relevant – and these assumptions are often as open to question as whether nonhumans have the relevant capacities.

So, whilst failed replication *might* be evidence about the reality of an effect, it might instead be evidence about these background assumptions. To put it another way, the failure of a direct replication attempt might be evidence that we have not performed a direct replication at all.[4]

Conversely, let us suppose instead that the replication succeeds, yielding results similar to those obtained in the original study. Once again, it is a familiar point that the epistemic significance of this result is limited. As mentioned above, results appearing to confirm the existence of complex capacities in nonhumans are often compatible with leaner interpretations – in many cases, because there may be systematic confounds. For instance, in many theory of mind experiments, it is argued that observable cues like line of gaze are systematic confounds, and that animals could use these cues to predict a target individual's behaviour, rather than ascribing any mental state (Penn & Povinelli, 2007). Obviously, a successful direct replication can do little to allay these sorts of worries about an experiment's validity. As Uljana Feest puts it, a direct replication can 'provide evidence for the existence of something, but it cannot say existence of what' (2019, p. 899). And conceptual replication has no answer to this, because to be confident that a conceptual replication has been performed requires confidence that the conceptual replication probes the same effect as the original study – but the individuation of the effect is precisely what is at issue (2019, p. 901).

But for similar reasons, to say that a successful direct replication provides evidence about the existence of *something*, whilst telling us nothing about *what*, is too simple. Drawing a conclusion about the reliability of the original study and the reality of its effect on the basis of the second study requires confidence that the latter was in fact a direct replication of the former, and that consequently the same effect was being probed in both cases. And as we have seen in the case of a failed replication, it is a non-trivial question whether this is the case in the situation described, in particular because open questions about the nature of the capacity under investigation might leave it unclear whether the experimental subjects, despite being conspecifics, are similar in relevant ways. These considerations are made salient by replication failure,

[4] On my view, whether one experiment directly replicates another turns on whether all factors relevant to producing the effect were in fact held fixed, which is independent of individual researchers' views about which factors are relevant. For this reason, I do not think that this kind of epistemic uncertainty can be eliminated by encouraging researchers to state their assumptions about which factors are relevant to reproducing the effect, though there are nevertheless good reasons for them to do so.

but remain in the background even where replications succeed. So, a successful replication provides evidence about the reality of an effect only given certain assumptions about the capacity in question, the experimental subjects and what factors are relevant.

Of course, one might point out that the success of the replication provides evidence about these background assumptions, as much as it provides evidence about the reality of the effect. That we found similar results in these two pools of experimental subjects suggests not only that there is an effect, but also that the subjects were similar in ways that are relevant. Supposing, for instance, that the first but not the second group of subjects were enculturated, we might take ourselves not only to have shown that there is an effect, but also that enculturation is not relevant when it comes to accounting for that effect. This is one respect in which it is too simple to say that the replication provides evidence for the existence of an effect whilst telling us nothing about the effect's nature: *if* it provides evidence for the existence of something, it certainly tells us something about what that thing is.

But another respect in which this is too simple is that the success of our replication may not provide univocal evidence about the existence of the effect. This is because the uncertainty surrounding our background assumptions will at least sometimes create space for disagreement about whether an experiment counts as a direct replication, and for alternative explanations of its results. For instance, suppose my theory about the capacity under investigation suggests that it should be present only in enculturated animals. Then I will be inclined to resist the claim that the second study counts as a direct replication of the first, and to offer an alternative explanation of its results: I might instead take it to indicate the existence of a distinct but superficially similar effect in the unenculturated subjects. How seriously I should be taken will depend upon the particulars of the case – including how my theory fits with other evidence and theory, its explanatory power, and so on. But given the theoretical openness characterising comparative cognition, it is likely that in at least some situations, there will be alternatives worth taking seriously. In such cases, the significance of successful replication will be underdetermined – in fact, it will again be underdetermined whether a study qualifies as a direct replication at all.

I have said that comparative cognition operates against a background of theoretical openness and uncertainty, in virtue of which its experimental results routinely underdetermine theory. Since replications are experiments, the same goes for them – with the additional complication that it will often be underdetermined whether one experiment 'counts' as a replication of another. As a result, neither successful replication nor failure to replicate provide clear evidence about the reliability of a previous study. This means that replications are poorly-placed to carry out the epistemic function sketched for them above, of evaluating the reliability of comparative cognition's methods. By extension, replication is poorly positioned in a demarcating role: since replications yield equivocal evidence about the reliability of comparative cognition's methods, treating the replicability of those methods as an indicator of their scientific *bona fides* will at best yield indeterminate results.

## 4. Progress

One might worry about the reliability of comparative cognition's methods for a number of reasons, including the possibility that social factors like publication bias, perverse incentives and questionable research practices may introduce bias into its published literature. These are serious concerns, but if what I have said here is correct, replications are likely to be of limited use in identifying or addressing them. Given the openness and uncertainty in comparative cognition's background, there may be any number of explanations for a replication's success or failure, including that we have not performed a replication at all. Fortunately, as Farrar and Ostojić (2019, pp. 13–15) highlight, there are a number of other approaches one might take to root out and address these influences, including 'open science' reforms. Less optimistically, what I have said might be taken to indicate a more fundamental concern: that even if comparative cognition's social organisation were epistemically optimal, we would be unable to draw meaningful conclusions from its experiments, or use replications to establish their reliability. So, one might be tempted to conclude that the prospects for making progress in comparative cognition are poor.

There is something right about this. In comparative cognition, experiments and replications are conducted against a backdrop of theoretical openness and uncertainty, and probe the nature of cognitive phenomena and their phylogenetic distribution simultaneously. Since both kinds of question are open and entangled with one another, there can be 'no direct application of some well-established theory to these questions' (Andrews, 2014, p. 15). So, by comparison with sciences whose background theory is better established, there are fewer constraints on the interpretation of experiments, meaning that we should not expect to be able to draw conclusions from experiment with any certainty.

Rather than being pessimistic about the prospects for progress in comparative cognition, though, I suggest that we revise our views about what scientific progress amounts to under conditions of uncertainty. In this vein, Kristin Andrews proposes that progress in comparative cognition is made not primarily by falsifying theories, but through a process of calibration: 'we start with a theory about the nature of some mental property, then we use that theory to make a considered judgment about whether some animal has that property, and use that judgment to empirically investigate the property. The results of that investigation might cause us to tweak our theory, our considered judgment, or both' (2014, p. 22). I agree with this picture – but would add that we often may not begin with one theory of a mental property, but with several candidate theories. Experimental results may not conclusively differentiate between theories, but may nevertheless help to calibrate them, by suggesting ways in which each should be 'tweaked'.

On this view, experimental results play an important role in the development of parallel theories, as well as in theory choice. I doubt that this is revisionary as a theoretical claim – but as Farrar and Ostojić (2019, p. 13) write, it may get lost in the 'day-to-day job of conducting empirical research'. Embracing this picture of experiment in practice might have some consequences for the way in which research in comparative cognition is organised.[5] Currently, researchers are incentivised to obtain highly constraining results and to draw concrete, eye-catching conclusions from them. Given the degree of openness in comparative cognition's background, it would be fruitful to place greater value on inconclusive results, and to incentivise a more wide-ranging exploration of the theories by which results might be explained.

A discussion of these theoretical options might provide the blueprint for follow-up studies probing particular aspects of the available theories (see Nosek & Errington, 2020). Follow-up studies will also be multiply interpretable, and may be viewed as replications on some theories but not others. But they will nevertheless be useful if they perform this same epistemic function – calibrating our theoretical options, by revealing ways in which the available theories might be tweaked. They may also narrow the range of available theories, where theories are unable to accommodate

[5] Thanks to two reviewers for suggesting this.

their results. With any luck, a process like this might generate a set of results compatible with only one theory. More likely is that inductive and abductive inferences will play a role in scientists' eventual theory choice – and correspondingly, about which of the follow up studies were replications. The interpretation of future results can then be constrained by this theory, reducing the underdetermination and uncertainty surrounding experiment and replication.

So although I have suggested that replications are poorly positioned to play the epistemic and demarcating roles sketched above, this is not to say that they should not be done – any more than the underdetermination characterising experiment in comparative cognition suggests we should abandon experiments. Rather, we should view replications and experiments as having similar function. The view I have been resisting casts experiment and replication in different roles: experiments lay bricks in a growing scientific edifice, and replications test the soundness of those bricks (Zwaan et al., 2018). Against this picture, it has been suggested that we should view scientific progress not on the model of building a wall, but of assembling a puzzle (Tullett & Vazier, 2018). This puzzle metaphor seems to fit better with the openness and uncertainty characteristic of comparative cognition, whose experiments seem to lack the finality of laying a brick. In this context, I suggest, experiment and replication play a similar role: both generate pieces of a puzzle which are, by themselves, difficult to make sense of. But taken collectively, and with the benefit of scientists' judgment, they provide clues about the sort of image we might be assembling, and what other pieces we ought to look for.[6]

[6] Thanks to Marta Halina, Ljerka Ostojić, Ivan Flis and an anonymous reviewer for this journal for providing detailed feedback on an earlier version of this paper.

References.

Andrews, K. (2014). *The Animal Mind: An Introduction to the Philosophy of Animal Cognition*. Routledge.

Derksen, M. (2019). Putting Popper to work. *Theory and Psychology*, *29*(4), 449–465. https://doi.org/10.1177/0959354319838343

Duhem, P. (1954). *The Aim and Structure of Physical Theory* (P. P. Wiener (ed.)). Princeton University Press.

Farrar, B., & Ostojic, L. (2019). The illusion of science in comparative cognition. *PsyArXiv*, *October 2*, 1–27.

Feest, U. (2019). Why replication is overrated. *Philosophy of Science*, *86*(5), 895–905. https://doi.org/10.1086/705451

Gallup, G. G. (1970). Chimpanzees: self-recognition. *Science*, *167*(3914), 86–87.

Heyes, C. M. (1994). Reflections on self-recognition in primates. *Animal Behaviour*, *47*(4), 909–919.

Heyes, C. M. (2018). Cognitive Gadgets. In *Cognitive Gadgets*. https://doi.org/10.4159/9780674985155

LeBel, E. P. (2017). *The Language of Science: A Primer [Blog Post]*. https://proveyourselfwrong.wordpress.com/2017/05/18/the-language-of-science-a-primer/

Lewens, T. (2015). *The Meaning of Science*. Penguin Random House.

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

Maes, E., Boddez, Y., Alfei, J. M., Krypotos, A. M., D'Hooge, R., De Houwer, J., & Beckers, T. (2016). The elusive nature of the blocking effect: 15 failures to replicate. *Journal of Experimental Psychology: General*, *145*(9), e49–e71. https://doi.org/10.1037/xge0000200

Maes, E., Krypotos, A. M., Boddez, Y., Matías, J., Palloni, A., D'Hooge, R., De Houwer, J., & Beckers, T. (2018). Failures to replicate blocking are surprising and informative-reply to Soto (2018). *Journal of Experimental Psychology: General*, *147*(4), 603–610. https://doi.org/10.1037/xge0000413

Nosek, B. A., & Errington, T. M. (2020). What is replication? *PLoS Biology*, *18*(3), 1–8. https://doi.org/10.1371/journal.pbio.3000691

Penn, D. C., & Povinelli, D. J. (2007). On the lack of evidence that non-human animals possess anything remotely resembling a 'theory of mind.' *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*, 731–744. https://doi.org/10.1098/rstb.2006.2023

Plotnik, J. M., de Waal, F. B. M., & Reiss, D. (2006). Self-recognition in an Asian elephant. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(45), 17053–17057.

Popper, K. (2002). *The Logic of Scientific Discovery*. Routledge.

Quine, W. V. O. (1958). Two Dogmas of Empiricism. In *From a Logical Point of View* (pp. 20–46). Harvard University Press.

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

Soto, F. A. (2018). Contemporary associative learning theory predicts failures to obtain blocking: Comment on Maes et al. (2016). *Journal of Experimental Psychology: General*, *147*(4), 597–602. https://doi.org/10.1037/xge0000341

Trestman, M. (2015). Clever Hans, Alex the Parrot, and Kanzi: What can exceptional animal learning teach us about human cognitive evolution? *Biological Theory*, *10*(1), 86–99. https://doi.org/10.1007/s13752-014-0199-2

Tullett, A. M., & Vazier, S. (2018). Scientific progress is like doing a puzzle, not building a wall. *Behavioral and Brain Sciences*, *41*(e120), 42–43.

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, *41*(e120), 1–13. https://doi.org/10.1017/S0140525X17001972