ABC 2021, 8(2):263-272
Animal Behavior and Cognition
DOI: https://doi.org/10.26451/abc.08.02.13.2021

Replications in Comparative Psychology

Marta Halina

Department of History and Philosophy of Science, University of Cambridge

Corresponding author (Email: mh801@cam.ac.uk)

Citation – Halina, M. (2021). Replications in comparative psychology. *Animal Behavior and Cognition, 8*(*2*), 263-272. https://doi.org/10.26451/abc.08.02.13.2021

Abstract – In order to assess the status of replications in comparative psychology, it is important to clarify what constitutes a replicated experiment. In this paper, I adopt the Resampling Account of replication recently advanced by the philosopher, Edouard Machery. I apply this account to a particular area of comparative psychology: nonhuman primate theory of mind research. Two key findings emerge from this analysis. First, under the account of replication advanced here, genuine replications are common in comparative psychology. Second, different types of replications offer different epistemic benefits to researchers. This second finding diverges from Machery's view, but provides insight into how the Resampling Account works in practice. I conclude that community-level change is needed in order to promote a wide range of replications and their associated diversity of epistemic benefits.

Keywords – Replication, Theory of mind, Comparative psychology, Great apes, the Resampling Account

In the last decade, efforts to replicate research have increased in a wide range of sciences, including medicine, psychology, and computer science (Bohannon, 2014). This initiative has emerged in part in response to a series of failed replication attempts and the growing awareness that certain standard scientific practices need to be evaluated and reformed (Romero, 2019). Thus far, comparative psychology has not been centrally involved in this so-called 'replication crisis' (but see Farrar & Ostojić, 2019; Farrar et al., 2020). In answer to the question, "Is there a 'replication crisis' in comparative psychology?" the psychologist, Michael Beran, writes, "Simply put, we do not know" (Beran, 2018, p. 1). In this paper, I aim to better understand the role of replications in comparative psychology. I do this by applying an account of replication recently advanced in philosophy of science (the Resampling Account) to an area of comparative psychology—namely, nonhuman primate theory of mind research. The Resampling Account provides a clear and principled account of replication, allowing us to identify those experiments in comparative psychology that should count as genuine replications. Crucially, I show that under this account genuine replications are common in comparative psychology. However, I argue that it is still important to distinguish between different types of replications as they support different inferences about the reliability of a study. This is a departure from Machery's account. Thus, the analysis provided here not only informs work in comparative psychology, but also results in a modification of the Resampling Account on the basis of its application to scientific practice. I conclude that community-level change is needed in order to support a wide range of replications and their associated diversity of epistemic benefits. In reply to the question, "Is there a 'replication' crisis in comparative psychology?" the answer then is "no and yes." "No" because much work in comparative psychology is dedicated to conducting replications; thus, some information about the reliability of studies in this field is already available. "Yes"

because major reforms are needed to ensure that replications are supported in a way that is on a par with novel studies—such support is currently lacking throughout the sciences.

Direct versus Conceptual Replications

It is common in psychology to distinguish between "direct replication" and "conceptual replication." Roughly, a direct replication tests the same hypothesis using very similar methods (treatment, setup, measurement, etc.) but different subject groups, while a conceptual replication tests the same hypothesis using different methods (Romero, 2019). Pashler and Harris (2012) argue that direct replications are epistemically critical for identifying false positives, but rare in psychology. They further argue that conceptual replications lack the epistemic benefits of direct replications and in fact likely lead to a greater proliferation of false positives.

Recently, the philosopher Edouard Machery has argued that direct and conceptual replications are both genuine replications with epistemic benefits (Machery, 2020). Specifically, Machery provides a general account of replication that he holds dissolves disagreements about the relative value of direct and conceptual replications. In the following section, I introduce Machery's account of replication before applying it to nonhuman primate theory of mind research.

The Resampling Account of Replication

In order to assess the status of replications in comparative psychology, it is important to have an account of what constitutes a replication. Machery (2020) advances a general account of replication, which he calls "the Resampling Account." According to this account:

> Experiment A replicates experiment B if and only if A consists of a sequence of events of the same type as B while resampling some of its experimental components in order to assess the reliability of the original experiment. (Machery, 2020, p. 556)

One experiment can never replicate another experiment exactly. However, according to the Resampling Account, two experiments should count as the same if the sequence of events that comprise them are tokens of the same type. A *token* event is a concrete instance of something, such as testing the treatment effects of a specific drug on a particular group of patients. Tokens are instantiations or exemplifications of *types*. The group of patients tested in a study is an instance of a broader category of individuals—perhaps all adult humans with asthma. Critically, types can be individuated in different ways. In comparative psychology, for example, one might ask whether a group of chimpanzees participating in a particular study is best understood as a token from the population of all chimpanzees, all adult chimpanzees, all adult captive chimpanzees, or all chimpanzees that have been enculturated by humans. Machery notes that type individuation is often left implicit in scientific work (2020, p. 556). However, type individuation is important because it indicates the scope of the hypothesis under investigation and what is relevant for testing it. For example, if a hypothesis concerns the type "nonhuman great apes", then researchers should ensure that the results hold for all three extant genera: *Pongo* (orangutans), *Gorilla* (gorillas) and *Pan* (chimpanzees and bonobos), whereas this would not be necessary if the hypothesis concerned the type "chimpanzees".

Individuating types is important for determining what counts as a replication. Under the Resampling Account, two experiments are identical for the purposes of replication if all of the events comprising the two experiments are tokens of the same type. If two experiments test different groups of participants—in one case, ten orangutans and the other ten chimpanzees—then, *ceteris paribus*, this is a genuine replication if the hypothesis refers to the type "nonhuman great apes". If the hypothesis instead concerns "chimpanzees" then this is an extension rather than a replication because the two studies sample different populations, rather than resampling the same population. As Machery notes, when it comes to

determining whether two experiments are identical in type, researchers "typically agree that some differences just do not matter, while others undoubtedly do" (2020, p. 550).

The above observations will not strike comparative psychologists as particularly surprising: it is common to think of the participants of an experiment as a sample or subset of a general population or type, which is then resampled for the purposes of a replication. If a replication is, as Romero (2019) notes, an "experiment whose design is identical to an original experiment's design in all factors that are supposedly causally responsible for the effect" (p. 2), then we can think of samples that come from the same type ("great apes") as identical in this sense. What is novel about the Resampling Account is that it highlights the fact that it is not just the participants of a study that can be treated as tokens of a population or type, subject to resampling, but other experimental components (such as treatment, measurement, and setting) as well. According to Machery, an experimental component is subject to resampling when researchers treat it as a "random factor" (2020). An experimental component is treated as a random factor when it 1) could have been different and 2) is meant to generalize to a broader class. For example, a study that tests the effects of a new cancer drug might rely on computed tomography (CT) scans to produce a measurement that serves as a dependent variable in the study. This is a random factor insofar as one could rely on a different form of measurement, such as the detection of biomarkers identified through liquid biopsy to determine the effects of the drug on treating cancer. Both of these measurements are meant to generalize to the broader class of measurements capable of detecting cancerous cells. Two studies that rely on these different forms of measurement are thus identical because the measurements are random factors sampled from the same population or type.[1]

Any experimental component (treatment, measurement, setting, participants) can be treated as a random factor. The result is that two experiments might look very different despite constituting a genuine replication. According to Machery, despite these differences, a replication of this kind can assess the reliability of an experiment in the same way that a direct replication would. Indeed, the Resampling Account holds that,

> A token experiment is reliable if and only if, if one repeatedly sampled new values for the experimental components that are treated as random factors… everything else being kept constant, the same experimental outcome would be found with high frequency (Machery, 2020, p. 555)

In the following section, I show how the Resampling Account applies to theory of mind research in comparative psychology. The application of this account has the surprising result that replications in this area of research are common. However, it also shows that some distinction between different types of replications should be maintained.

Replications in Comparative Psychology: Theory of Mind Research

Theory of mind (ToM) is the ability to attribute mental states to other agents. It is what we do when we predict and explain the behavior of others by appealing to their beliefs, desires, intentions, and perceptions, rather than just their observable behavior. ToM is thought to be ubiquitous in adult human life and underlie other cognitive abilities, such as empathy and self-awareness. Many of these abilities have long been held to be uniquely human; thus, discovering whether nonhuman animals mindread could have radical consequences on how we understand humans and other animals (Krupenye & Call, 2019).

The Resampling Account fits ToM research in comparative psychology well. As we will see below, researchers working in this area typically treat several experimental components as random factors. Broadly, ToM can be understood as the ability to produce and use information about the cognitive state of another agent (see Halina, 2015). Researchers working in this area generally treat as distinct the abilities to produce and use information about different cognitive states, such as goals, intentions,

[1] The term "random" should be construed broadly here to include convenience samples (see Machery, 2020, p. 553).

perceptions, and beliefs (Call & Tomasello, 2008). For the purposes of this discussion, I will focus on research aimed at investigating great apes' understanding of perceptual states. Researchers working in this area examine whether nonhuman primates, such as chimpanzees, understand what other agents can see. Following Lurz et al. (2018), we can call this the "seeing hypothesis."

As noted above, a random factor is an experimental component that could have been different: in a given experiment, such a factor represents a sample from a broader population or type and researchers aim to generalize from the sample under investigation to this type. In contrast, fixed factors "exhaust the relevant population" (Machery, 2020, p. 552). In this case, the conclusions drawn by researchers extend only to those values of the experimental component that were tested in a given experiment. Applying this distinction to studies aimed at investigating the seeing hypothesis reveals that researchers typically treat participants, treatment, and measurement as random factors. This is evidenced by the scope of the claims or conclusions drawn from these studies. Consider the titles of some key research papers in this area: "Chimpanzees know what conspecifics do and do not see" (Hare et al., 2000), "Chimpanzees really know what others can see in a competitive situation" (Bräuer et al., 2007), "Do chimpanzees know what each other see? A closer look" (D'Arcy & Povinelli, 2002). These titles indicate the target of analysis: whether chimpanzees are sensitive to (or "know" or "understand") whether another agent can or cannot see a particular object (also known as level 1 visual perspective taking). Accordingly, Hare et al. (2000) report finding that "chimpanzees know what conspecifics can and cannot see, and that they use this information flexibly in certain competitive situations" (p. 784). Similarly, Hostetter et al. (2007) conclude: "it appears that chimpanzees do understand the difference between someone who can see them and someone who cannot" (p. 62). Melis et al. (2006) also write that their results, combined with similar studies, "demonstrate that chimpanzees have some understanding of what others can see" (p. 161). Importantly, these studies do not claim to have exhaustively observed all chimpanzees (participants), all states of seeing and not seeing an object (treatment), and all states of chimpanzees' knowledge of the perceptual states of others (measurement). Instead, researchers are treating these three experimental components as random factors. They extend their claims from the samples observed in a particular experiment to these populations. They do not limit their claims, for instance, to a group of chimpanzees housed at the University of Louisiana at Lafayette (D'Arcy & Povinelli, 2002), but extend them to the species *Pan troglodytes*. They do not limit their claims to situations involving opaque and transparent barriers, but extend them to the broader type of situations in which an agent can or cannot see something (Hare et al., 2000). Finally, researchers do not limit their claims to one set of measurements, such as reaching for food (Bräuer et al., 2007), but extend them to the broader population of measurements that determine whether subjects produce and use information about another agent's perceptual state.

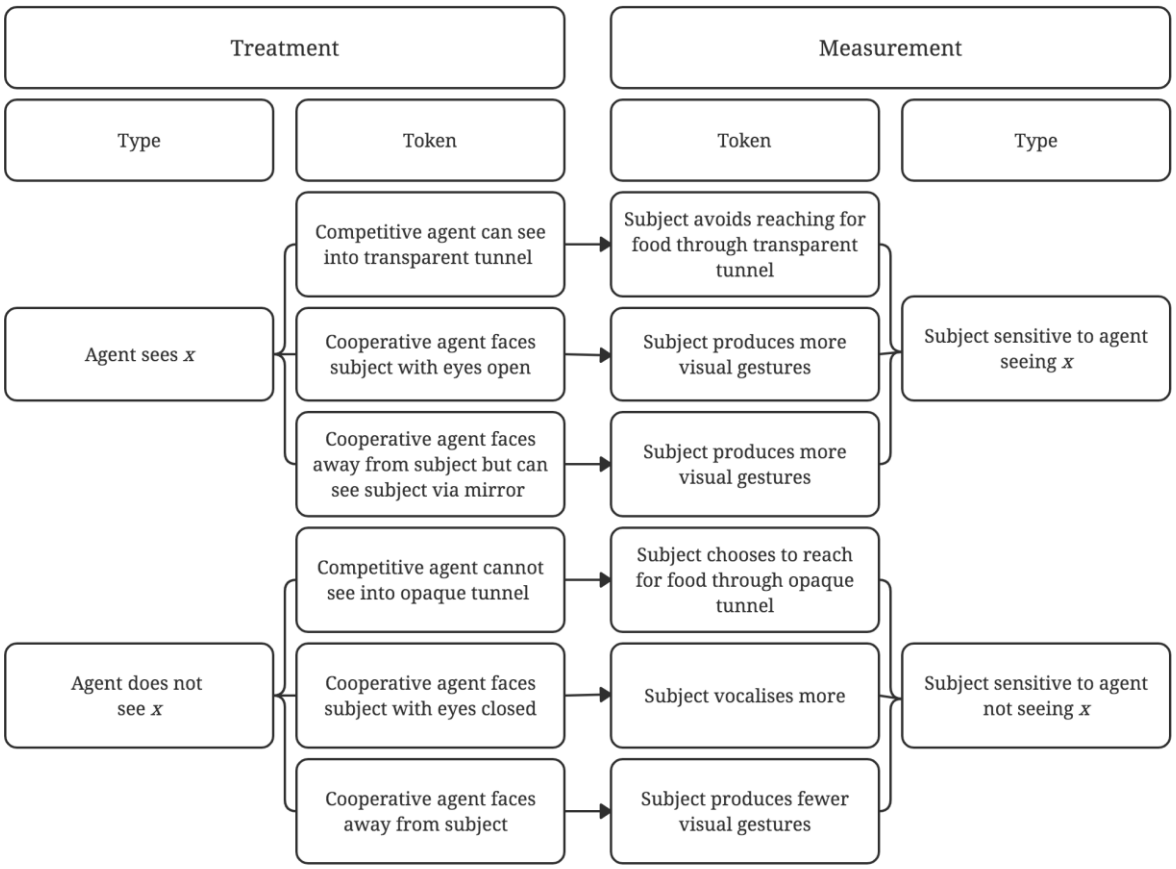
Researchers explicitly identify some of the studies aimed at testing the seeing hypothesis in chimpanzees as replications.[2] Several of these replications resemble very closely the original study. Karin-D'Arcy and Povinelli (2002), for example, replicate the experiments of Hare et al. (2000) using the same methods on a different group of chimpanzees. Bräuer et al. (2007) replicate Hare et al. (2000) with "a new group of chimpanzees" and "slightly different methods" (p. 441). Kaminski et al. (2004) replicate a combination of the studies performed by Call and Tomasello (1994) and Povinelli and Eddy (1996a), writing, "We used an experimental paradigm previously used with two orangutans (Call and Tomasello 1994) to administer some of the conditions used by Povinelli and Eddy (1996a)" (p. 217). Although many of these replications follow closely the methods of the original study, this is not always the case. In those cases, explicitly identified as "replications", however, researchers often note the differences between the new and original study. Karin-D'Arcy and Povinelli (2002), for example, note that their testing arena was 2.6 x 1.8 m, which was "slightly smaller than the 3 x 3 m testing arena used by Hare et al." (p. 31). Under the Resampling Account, the above studies count as replications. This is unsurprising, however, given that these studies follow closely the methods of the original experiment, while resampling the subjects under study, and thus fall under standard accounts of "direct replication." Crucially, the Resampling Account also counts many other studies in this domain as genuine replications, provided they are

[2] I use the term "replication" to indicate a replication attempt, rather than a successful replication.

resampling random factors from the same population or type. Such studies are identical for the purposes of replication as noted above. This includes studies that differ considerably in the particular treatments and measurements they employ, as depicted in Figure 1. Such studies are numerous in ToM research (see Call & Tomasello, 2008; Krupenye & Call, 2019, for reviews).

Figure 1

An Example of Type and Token Treatments and Measurements in ToM Research



Note. The descriptions of treatments and measurements are based on the studies of Melis et al., 2006 (transparent versus opaque tunnel), Hostetter et al., 2007 (eyes open versus eyes closed), and Lurz et al., 2018 (mirror versus no mirror).

Studies such as those depicted in Figure 1 have not previously been identified as replications. The reason for this may be that researchers typically do not acknowledge treatments and measurements as random factors (Machery, 2020). However, based on the claims made in these and related studies, the treatments and measurements are being treated as random, rather than fixed, factors. In the same way that such studies regularly make claims about the type "chimpanzees", they also make claims about the type "perceptual states of another agent" and the type "sensitivity to the perceptual states of another agent."

Why are replications common in this area of research? One reason is that early work testing the seeing hypothesis produced mixed results. While some studies produced negative results (see Bräuer et al., 2007 for a review), others produced positive results (e.g., Hare et al., 2000, 2001). In the last twenty years, a consensus that chimpanzees produce and use information about the perceptual states of other agents has emerged (Call & Tomasello, 2008; Krupenye & Call, 2019). This consensus has been reached in part because researchers have provided compelling alternative explanations for the available negative results. For example, Hare et al. (2000) attribute the negative results of Povinelli and Eddy (1996b) and

others to the cooperative (as opposed to competitive) nature of the tasks, stating that "chimpanzees' most sophisticated social-cognitive abilities may emerge only in the more natural situation of food competition with conspecifics" (p. 783). Kaminski et al. (2004) attribute the negative results of Povinelli and Eddy (1996a) to a lack of ecological validity, noting that, "chimpanzees needed to be trained over hundreds of trials to participate in this experiment meaningfully" (p. 217). They proceed to use a method that does not require training, but instead relies on "apes' natural tendency to beg for food that is beyond their reach" (Kaminski et al., 2004, p. 217). Finally, Bräuer et al. (2007) attribute the negative results of Karin-D'Arcy and Povinelli (2002) to the experimental setting, arguing that the distance between food items (1.25 m) was too small. They proceed to replicate the study with a larger distance between food items (2 m) and obtain positive results.

If the Resampling Account is correct, and ToM researchers treat participants, treatment, and measurement as random factors, then this means that many studies aimed at testing the seeing hypothesis are genuine replications, regardless of how closely they follow the methods of an original study. Moreover, the concern expressed by Pashler and Harris (2012) that conceptual replications exacerbate false positives would not apply because, under the Resampling Account, the distinction between direct and conceptual replications dissolves. As Machery writes, the account leads one to "abandon the usual distinction between direct and conceptual replication and thus any attempt to establish the epistemic superiority of one of them" (2020, p. 565). Whether an experiment resamples participants (as in a direct replication) or treatment or measurement (as in a conceptual replication) is not important. What is important is that an experiment resamples random factors from the same type.

Although I generally agree with the above conclusion and think it is a virtue of the Resampling Account that it dissolves the distinction between direct and conceptual replications, I also think that a closer look at ToM research reveals that there is a distinction between different types of replications that should be maintained. Crucially, two studies that are very different in their methods due to the resampling of multiple random factors from broad or coarse-grained types are likely to vary in their auxiliary assumptions. This makes it difficult to determine whether a failed replication indicates a lack of reliability or is instead a result of a relevant difference in auxiliary assumptions. To see this, consider a replication that aims to match the original study very closely. Karin-D'Arcy and Povinelli (2002) aim to replicate a chimpanzee ToM experiment (reported in Hare et al., 2000) involving placing a piece of food on top of a tire and another piece of food inside the tire. This replication resampled the population of chimpanzees, but the authors attempted to provide this new group of chimpanzees with the same experiences as the original group of participants, as well as reproduce the other experimental components as closely as possible. Despite this, however, they encountered problems in conducting the replication. They write,

> In attempting to replicate this experiment we encountered insurmountable procedural difficulties in following the published method… Correspondence with the senior author of the original study revealed that they encountered similar problems that were solved by an unmentioned procedural change (Karin-D'Arcy & Povinelli, 2002, p. 30)

Failing to replicate the original study in this case led the authors to identify an important change to the published methods. If, on the other hand, the authors had resampled the treatment and measurement, in addition to the participants, then the cause of the failed replication would be more difficult to identify. In this latter case, the cause might instead be best attributed not to a lack of experimental reliability or miscommunication of methods, but to a change in the procedure or one or more auxiliary assumptions. Indeed, historically in ToM research, failed replications that have diverged significantly in one or more experimental components from the original experiment have been explained in this way. This feature of experimental work is made explicit in a recent special issue in *Cognitive Development* on replications of implicit false belief. As Dörrenberg et al. (2018) write, regarding one replication study, "given that this was a conceptual rather than direct replication, with many differences between original and replication task, [the absence of evidence of replicability] by itself cannot be considered evidence for the absence of robust replicability" (p. 26). The editors of this issue also

explicitly reference the Duhem-Quine thesis and the many implicit auxiliary assumptions that feed into the construction and implementation of experiments (Sabbagh & Paulus, 2018). Thus, although the Resampling Account dissolves the distinction between a direct and conceptual replication, it seems important to retain a distinction between a replication that resamples one experimental component versus many experimental components at once.[3] A failed replication in the latter case may indicate little about the reliability of the original study, given the many alternative explanations available for the negative result.

Given the above, one might share the concern expressed by Pashler and Harris (2012) regarding conceptual replications: if researchers regularly "explain away" failed replications, particularly when those replications differ widely in their token methods or auxiliary assumptions, then perhaps these replications cause more epistemic harm (such as an accumulation of false positives) than benefits. In the case of nonhuman primate ToM research, I believe such a picture oversimplifies the nature of experimental practice. Instead, what we find is that researchers obtain new information from failed replications and adjust future theorising and experimental design in response to this information. Such negative results thus provide epistemic benefits in helping researchers improve their understanding of the phenomenon under investigation and the best experimental approach for investigating it. This is in contrast to Pashler and Harris's view. They write regarding failed conceptual replications that, "Rarely, it seems to us, would the investigators themselves believe they have learned much of anything" (2012, p. 533). Although it might be true that researchers do not believe they have learned much about the reliability of a study in this case, they do typically take themselves to have learned something about the target phenomenon or experimental approach in question, at least in ToM research in comparative psychology.

I have suggested that it is important to maintain a distinction between a replication that resamples one versus multiple experimental components. This may seem in tension with Machery's account. As noted, the Resampling Account holds that two experiments are identical for the purposes of replication regardless of how many experimental components are resampled provided they are resampled from the same types as the original study. In a case of multiple factors being resampled, Machery writes, "while the follow-up experiment would be superficially different, it would be the same experiment at a deeper level" (2020, p. 558). Introducing a distinction between replications that resample one versus several random factors seems in tension with the idea that these studies are identical for the purposes of replication. It also sounds like we are smuggling back in the distinction between direct and conceptual replications. There are two responses to make here. First, Machery's main complaint against the direct-conceptual replication distinction is that it is not grounded in a principled account of replication. In contrast, any distinctions that emerge out of the Resampling Account (such as the one proposed here) do not face this problem. Second, Machery distinguishes among replications that resample different experimental components, labelling these "treatment replication" or "measurement replication" for instance (2020, p. 562). Thus, it seems like a small step to add to this taxonomy the idea of a replication that involves the resampling of one versus several random factors. Machery then need not resist this addition to his account. Not all replications are on a par with respect to their ability to assess reliability under the Resampling Account. When a replication resamples several random factors (particularly, when these samples are taken from broad or diverse populations), and this replication fails, this is not a clear indicator of poor reliability. Instead, differences in auxiliary assumptions may plausibly explain this discrepancy in results. In contrast, when a replication resamples one random factor (particularly when this sample is drawn from a narrow or homogenous population), and this replication fails, this is a clearer indicator of poor reliability, given the few alternative explanations available. [4]

[3] This is particularly important if the population from which an experimental component is resampled is very broad or coarse-grained, as two samples (and their corresponding auxiliary assumptions) drawn from such a population may differ considerably while still counting as identical under the Resampling Account.

[4] Machery is aware of the claim that "the failure of a conceptual replication is said to be uninformative because it could result from relevant differences between the original experiment and its conceptual replication" (2020, p. 546), but does not discuss this issue with respect to the Resampling Account.

Applying the Resampling Account to ToM research in comparative psychology reveals that replications in this field are common. Some of these replications resample from the population of participants while keeping the value of other factors at the level of the original experiment, while others resample multiple random factors at once: participants, treatment, measurement, and setting. Furthermore, many of these replications have produced successful results. This suggests that studies investigating whether chimpanzees produce and use information about the perceptual states of others are reliable. If these studies were unreliable, then one would not expect the same result to repeatedly obtain when new values of the experimental components were sampled.[5] Unfortunately, two complications prevent us from concluding that these studies are reliable without further analysis.

First, as argued above, the picture is complicated by the nature of failed replications. An important difference between a replication that resamples one versus multiple random factors is that the latter provides fewer constraints on how to interpret negative results. Thus, in practice, a failed replication of this latter kind does not typically threaten the reliability of an original study (although it may have other epistemic benefits as discussed above). Thus, in an area like ToM research, where multiple random factors are often resampled from coarse-grained types, numerous successful replications alone are not a good indicator of reliability. The number of failed replications and how they are interpreted is important to consider as well. Such an analysis of failed replications in ToM research goes beyond the scope of this paper, but is an important avenue for further research. In the meantime, the Resampling Account can help us identify replications in this area and provide a better understanding of how experiments contribute to reliability. This account also draws attention to the crucial fact that the resampling of participants is only one of several ways of establishing experimental reliability when multiple experimental components are treated as random factors as in the case of ToM research.

This leads us to a final complication. Research on ToM in comparative psychology has advanced in large part due to the many replications conducted in this domain. Failed replications have played a particularly important role in these advances. Whether all failed replications in this domain are published or made available, however, is unknown. Given the numerous disincentives and obstacles in place across the sciences for both conducting replications and publishing negative results, it is likely that many studies of this kind have ended up in the file drawer (Romero, 2017; Rosenthal, 1979). This is a problem that needs to be urgently addressed in comparative psychology, as well as other sciences. Replications are needed not only to assess the reliability of experiments, but also to help researchers identify and understand the auxiliary assumptions relevant for investigating a target phenomenon.

A central focus of the replication crisis has been to seek out the "questionable research practices" that might explain failed replications (Bohannon, 2014). Although it is important to identify such practices (Simmons et al., 2011), it is also crucial to emphasise that good science will include failed replications—such failed replications should be viewed as sources of knowledge: epistemic benefits, not failures. Comparative psychologists should join other scientists in changing the structures and norms that discourage and stigmatise replications. Daniel Kahneman has urged researchers to "get in front of the problem by replicating their own work" (Bohannon, 2014, p. 789). However, community-level change is needed.

Conclusion

In this paper, I have adopted Machery's Resampling Account of replication in order to better understand the role of replications in comparative psychology. This analysis has revealed that in the domain of theory of mind research, comparative psychologists treat multiple experimental components as random factors and have conducted numerous studies that resample one or more of these factors. Moreover, many of these replications have produced positive results, suggesting that the ToM experiment

[5] Note that reliability should be distinguished here from the issue of controlling for directional error. If, for instance, the seeing hypothesis makes the same predictions as a different hypothesis—such as one that posits sensitivity to line of gaze (Lurz et al., 2018)—then this is a problem for the seeing hypothesis, despite the reliability of these experiments.

examined here has high reliability. However, this suggestion must be treated with caution, as the exact number and nature of failed replications in this area is unknown. Although we have a good understanding of those replications that have been published, it is plausible that there are failed replications that remain unpublished, given the existing barriers to publishing work of this kind. Machery's Resampling Account provides a clear and principled framework for identifying and analysing replications in comparative psychology. Applying this framework reveals the epistemic benefits generated by successful and failed replications, which in turn may help reform those practices that currently obstruct such work.

Acknowledgements

Many thanks to Ali Boyle, Benjamin Farrar, and Ljerka Ostojić for early discussions of this paper. I am also grateful to an excellent group of PhD students (David Harrison, Richard Ngo, Daniel Ott, Erlend Owesen, Victor Parchment and Konstantinos Voudouris) who not only provided valuable feedback on this paper, but have also maintained a positive and supportive environment throughout this challenging year. Finally, special thanks to Edward Legg, Mike Dacey and an anonymous reviewer for their thoughtful and detailed comments.

References

Beran, M. (2018). Replication and pre-registration in comparative psychology. *International Journal of Comparative Psychology*, *31*, 1–8.

Bohannon, J. (2014). Replication effort provokes praise—and 'bullying' charges. *Science*, *344*(6186), 788–789.

Bräuer, J., Call, J., & Tomasello, M. (2007). Chimpanzees really know what others can see in a competitive situation. *Animal Cognition*, *10*(4), 439–448.

Call, J., & Tomasello, M. (1994). Production and comprehension of referential pointing by orangutans (*Pongo pygmaeus*). *Journal of Comparative Psychology, 108*(4), 307–317.

Call, J., & Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Sciences*, *12*(5), 187-192.

Dörrenberg, S., Rakoczy, H., & Liszkowski, U. (2018). How (not) to measure infant theory of mind: Testing the replicability and validity of four non-verbal measures. *Cognitive Development*, *46*, 12–30.

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.

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

Halina, M. (2015). There is no special problem of mindreading in nonhuman animals. *Philosophy of Science*, *82*, 473–490.

Hare, B., Call, J., Agnetta, B. & Tomasello, M. (2000). Chimpanzees know what conspecifics do and do not see. *Animal Behaviour*, *59*, 771–785.

Hare, B., Call, J., & Tomasello, M. (2001). Do chimpanzees know what conspecifics know? *Animal Behaviour*, *61*(1), 139–151.

Hostetter, A. B., Russell, J. L., Freeman, H., & Hopkins, W. D. (2007). Now you see me, now you don't: Evidence that chimpanzees understand the role of the eyes in attention. *Animal Cognition*, *10*(1), 55–62.

Kaminski, J., Call, J., & Tomasello, M. (2004). Body orientation and face orientation: Two factors controlling apes' begging behavior from humans. *Animal Cognition*, *7*(4), 216–223.

Karin-D'Arcy, R. M., & Povinelli, D. J. (2002). Do chimpanzees know what each other see? A closer look. *International Journal of Comparative Psychology*, *15*(1), 21–54.

Krupenye, C., & Call, J. (2019). Theory of mind in animals: Current and future directions. *Wiley Interdisciplinary Reviews: Cognitive Science*, *10*(6), e1503.

Lurz, R., Krachun, C., Mahovetz, L., Wilson, M. J., & Hopkins, W. (2018). Chimpanzees gesture to humans in mirrors: Using reflection to dissociate seeing from line of gaze. *Animal Behaviour*, *135*, 239–249.

Machery, E. (2020). What is a replication? *Philosophy of Science*. *87*, 545–567.

Melis, A. P., Call, J., & Tomasello, M. (2006). Chimpanzees (*Pan troglodytes*) conceal visual and auditory information from others. *Journal of Comparative Psychology*, *120*(2), 154-162.

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, *7*(6), 531–536.

Povinelli, D. J., & Eddy, T. J. (1996a). What young chimpanzees know about seeing. *Monographs of the Society for Research in Child Development*, *61*(3), v-vi, 1–191.

Povinelli, D. J., & Eddy, T. J. (1996b). Chimpanzees: Joint visual attention. *Psychological Science*, *7*(3), 129–135.

Romero, F. (2017). Novelty versus replicability: Virtues and vices in the reward system of science. *Philosophy of Science*, *84*(5), 1031–1043.

Romero, F. (2019). Philosophy of science and the replicability crisis. *Philosophy Compass*, *14*(11), e12633.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin*, *86*, 638–641.

Sabbagh, M. A., & Paulus, M. (2018). Replication studies of implicit false belief with infants and toddlers. *Cognitive Development*, *46*, 1–3.

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