Doing Reliable Research in Comparative Psychology: Challenges and Proposals for Improvement

Emma C. Tecwyn

Department of Psychology, Birmingham City University, Birmingham, UK

emma.tecwyn@bcu.ac.uk

Unlike some other areas of psychology that have experienced a 'reproducibility crisis', the extent to which research findings in comparative psychology are reliable is only just beginning to come under the spotlight. I outline what is known about where we as a field stand in terms of the reliability of our findings, and highlight some characteristic features of our research that give may cause for concern, focusing primarily on experimental comparative cognition. I then discuss ways that we as individual researchers and a wider community can take steps to improve our current practices (and in some cases already are), as well as highlighting the crucial role institutions and gatekeepers have to play in effecting change. By tackling potential issues head on, the field of comparative psychology can have more confidence that our research findings and the resultant claims we make about animal behaviour and cognition are reliable.

Keywords: comparative psychology; comparative cognition; reliability; reproducibility; open science

Comparative psychology investigates, via empirical studies, how animals including humans acquire, process and use information (Call et al., 2017; Shettleworth, 2010). The comparative approach enables us to better understand the behaviour and cognition of individual animal species, as well as to identify similarities and differences between species, and thus make inferences about their adaptive significance and the selection pressures that may have driven the evolution of specific processes (Smith et al., 2018). The field encompasses diverse subtopics, and uses a variety of methodological approaches to study wild and captive individuals of an increasingly wide array of species (though it has been argued that there is still a need to strategically diversify, in order to increase the power and resolution of hypotheses being tested, and enable new questions to be addressed, Call et al., 2017). Observational and experimental studies are conducted in the field and laboratories, and increasingly in zoos and sanctuaries (Hopper, 2017). All of these approaches have a crucial and complementary role to play, with studies of wild individuals in the field allowing animals to demonstrate their natural behaviour/cognition in the environment in which it evolved; and experiments with captive individuals enabling controlled investigation of the mechanisms underpinning naturally observed behaviour (e.g., Boesch 2007; 2008; Kamil, 1987; Snowdon & Burghardt, 2017; Tomasello & Call, 2008). To illustrate the current diversity of comparative psychology research, in the past five years this journal has published studies on: embracing in geladas (Pallante et al., 2019); physical cognition in giraffes (Caicoya et al., 2019), manual lateralisation in kangaroos (Giljov et al., 2017); self-toxicity-recognition in snakes (Mori & Burghardt, 2017); and the development of ethograms for octopods (Mather & Alupay, 2016), to name but a few.

Being such a diverse and complex field of research, comparative psychology has long faced challenges and crises of confidence, many of which are ongoing, including a lack of theory (Boyle, 2021; Hodos & Campbell, 1969; Vonk & Shackleford, 2012; Wynne & Bolhuis, 2008), and agreeing on what the central aims of the field should be (Burghardt, 2013; Hirsch, 1987). A relatively new challenge for our field—and the one that I will focus on here—concerns the extent to which the research findings we produce and the claims we make are reliable (objectively credible due to formal correctness and reproducibility; Stracke, 2020). This challenge is by no means unique to the field of comparative psychology, but due to some characteristic features of our research, it may be particularly difficult to ascertain the reliability of our existing research record, as well as to address these concerns going forward. However, comparative psychology researchers have recently begun to shine a spotlight on potential reliability

issues, and to propose ways in which we as a field can make progress.

Given the breadth and diversity of comparative psychology, the challenges regarding reliability differ to some extent between subfields, and there is no 'one size fits all' approach in terms of potential solutions. Here, I will focus primarily on one specific subfield and methodological approach within comparative psychology—experimental comparative cognition—which aims to compare cognitive processes across species to identify similarities and differences. As well as better understanding the cognition of different species (including humans), this subfield aims to uncover how cognitive abilities evolve, and the processes that drive this evolution (MacLean et al., 2014; Krasheninnikova et al., 2020; Schubiger et al., 2020). A focus on comparative cognition reflects my interest in this area, but also, as I will set out, the nature of research in this particular subfield of comparative psychology raises specific challenges in relation to reliability. While not all agree with the growing focus on comparative cognition, and have particular concerns regarding an anthropocentric focus (e.g., Hodos & Campbell, 1969; Hirsch, 1987; Vonk & Shackleford, 2012; Burghardt, 2013), learning and cognition have dominated the field since the late 20th century (Call et al., 2017), and articles related to learning and cognition have increased over time in this journal (Burghardt, 2013). Given that comparative cognition continues to be a very active area of research, there is value in highlighting reliability concerns in this subfield. However, many of the issues I discuss also apply to comparative psychology more broadly, so although I will give specific examples from experimental comparative cognition, where appropriate I make reference to other subfields and methodological approaches, and highlight how challenges and potential solutions may or may not apply to these. I will also focus primarily on null hypothesis significance testing, given that this approach to statistical analysis remains common at present in comparative psychology.

In places I draw parallels with cognitive development research—again, this reflects my interest in this field, but also because doing research with young children (especially pre-verbal children) involves many similar methodological challenges to doing research with nonhuman animals, and the field of cognitive development, and developmental psychology more broadly, interact productively with comparative psychology (Beck, 2017; Call et al., 2017). Furthermore, cognitive development researchers—especially those doing research with infants—have recently identified issues with the replicability of some findings (e.g., Kampis et al., 2020; Kulke et al., 2018a; 2018b) and have already begun to take steps towards improving the reliability of their research (e.g., Byers-Heinlein et al., 2021), and thus there may be valuable lessons to be learned for those of us working with nonhuman animals.

My aim is to bring attention to potential reliability issues in comparative psychology and highlight ways in which we might improve our practices to increase the replicability of our findings. I first provide a brief background to concerns regarding the reliability of research findings in psychology more broadly, before outlining where we as a field stand, highlighting some features of comparative psychology research (and experimental comparative cognition in particular) that may give cause for concern. I then discuss ways in which we as comparative psychology researchers can improve our practices, as well as the role that institutions and gatekeepers have to play in facilitating structural change.

How reliable are findings in comparative psychology?

Scientific progress depends on data being reliable. Over the past decade, the realisation that many original findings in psychology do not replicate (e.g., Camerer et al., 2018; Open Science Collaboration, 2015) has made clear that existing research practices can undermine scientific progress, leading to the questioning of psychology as a trustworthy science. Several factors have been identified that have likely contributed to this 'reproducibility crisis' (Baker, 2016). These include researcher 'degrees of freedom'—the arbitrary decisions made by researchers throughout data collection and analysis (Gelman & Loken, 2013), which can lead to unethical or questionable research practices (QRPs). QRPs can occur throughout the research process (e.g., Ioannidis, 2005; Munafo et al., 2017; Simmons et al., 2011). For example, researcher bias can occur during the specification of hypotheses, as well as during the interpretation of results in the form selective reporting of tests that 'worked' (cherry picking), p-hacking (e.g., deciding whether to exclude data after analysis; Simmons et al., 2011) and HARKing (hypothesising after results are known, in order to fit the data; Kerr, 1998). Finally, problematic incentive structures in terms of journal publications and their associated metrics (e.g., impact factors, h-index) lead to publication bias, wherein 'positive' novel findings (typically where $p < .05$) are more likely to be published than null findings,

'messy' data, or replications (e.g., Edwards & Roy, 2017; Lawrence, 2016; Munafo et al., 2017; Nosek et al., 2012; Scheel et al., 2020; Smaldino & McElreath, 2016).

To date, the field of comparative psychology has not come under the spotlight in terms of issues with the reliability of the field's findings (Halina, 2021). For some other areas of psychology there is evidence that even many 'established' findings do not replicate. For example, of 100 psychology studies, only 36% of attempted replications produced a significant result, compared with 97% of the original studies (Open Science Collaboration, 2015). In contrast, we do not currently know the extent to which we can be confident about previous findings in comparative psychology, given the scarcity of direct replications (Beran, 2018). In the closely related field of behavioural ecology, Kelly (2006) reported that although partial or conceptual replications made up around 30% of articles in three top journals, there were no exact replications. More recently, the same author reported that only 0.023% of articles within the ecology and evolution literature claimed to be 'true' replication studies (Kelly, 2019). However, there is some evidence of replication issues in animal behaviour. For example, classic research on female choice in zebra finches purported to show a preference for males with red leg bands over those with green leg bands (e.g., Burley, 1981), but a recent article that included multiple experiments and a meta-analysis of previous studies found no effects of leg band colour on reproductive success in zebra finches (Wang et al., 2018). Farrar and colleagues (2020) used simulation to demonstrate that just-significant findings (those where the p-value falls between 0.01 and 0.05), which are prevalent in the comparative cognition literature as a result of publication bias, are likely to replicate around 50% of the time.

Despite the scarcity of evidence regarding the reliability of comparative psychology findings, researchers in ecology and evolution (fields closely related to ours) report engaging in QRPs at comparable levels to other areas of psychology, with 51% of respondents reporting engaging in HARKing, 42% collecting additional data after conducting statistical analyses to check for significance, and 64% selectively reporting 'positive' results and not including non-significant findings (Fraser et al., 2018). On a more positive note, a recent survey of 63 papers in the animal physical cognition literature showed that some papers (10-17%) did report negative findings, and a further 24-46% made inconclusive claims, demonstrating that 'negative' findings can and have been published (Farrar et al., 2021).

Why might we be particularly concerned about the reliability of research findings in comparative psychology—and comparative cognition in particular? In addition to the unique challenges posed by working with animals and measuring their behaviour (Brecht et al., 2021), the 'standard approach' in experimental comparative cognition for many years has been something along the following lines: a research group comes up with a question/hypothesis about a particular cognitive ability in their study species. They develop a novel apparatus to probe this cognitive skill and, without assessing the validity of the task, use it to test their animals (usually a small number of individuals), which are housed at a particular site and have a specific test history. The animals' behavioural responses—often a choice between two options—are recorded over a number of trials. The overall performance of the group is aggregated, and if the average performance of the group differs from chance level, or differs between critical conditions, this is taken as evidence for the presence of the cognitive ability under investigation in this species (e.g., Thornton & Lukas, 2012; Shaw & Schmelz, 2017). If overall performance does not differ from chance, typically no firm conclusions are drawn; instead *post hoc* explanations for the species' 'failure' are discussed (e.g., task design issues; performance limited by non-cognitive factors, etc; Farrar & Ostojic, 2019). Another group then tests for this same ability in a different species, but tweaks the apparatus and testing procedure to suit their testing site/individuals, or even uses a different methodological approach entirely. Multiple studies of this nature are then compared, and conclusions are drawn regarding which species do or do not possess the cognitive ability in question (Beran, 2018; Farrar & Ostojic, 2019; Farrar et al., 2020; Krashenninikova et al, 2020).

While this description clearly lacks nuance and by no means applies to all comparative cognition studies (for example, unlike many other psychology disciplines, lots of studies report and discuss individual- as well as group-level performance, thus providing a richer dataset that provides a better indication of the robustness of group-level effects; e.g., Thronton & Lukas, 2012), it serves to highlight some features of our research process that might mean there is cause for concern regarding the replicability of the field's findings. In what follows, I discuss some of these features and the associated challenges in more detail, including whether and to what extent they apply to other areas of comparative psychology.

Field-specific challenges

Hypothesis formulation

Farrar & Ostojic (2019) express concerns that the current standard approach may not be appropriate for addressing whether species possess certain cognitive abilities. They highlight the issue of a directional bias in hypothesis formulation, such that if a species displays behaviour X, then this is taken as evidence for cognitive skill Y. Thus, the null hypothesis is that the species does not display behaviour X; however, this outcome is rarely, if ever, used to reject alternative hypothesis; instead, there is discussion of the difficulty of interpretation and/or *post hoc* consideration of potential explanations for null results. In sum, it is rare in the field that hypotheses are set up to test competing theories, and as a result, only positive results mean anything (Farrar & Ostojic, 2019). Relatedly, others have expressed concerns around the lack of a theoretical framework in comparative psychology more broadly, which has resulted in the exploration of behaviour without any clear focus (e.g., Hodos & Campbell, 1969; Vonk & Shackleford, 2012).

Data collection and analysis

Small sample sizes are common in comparative psychology, being constrained by a research group's access to individuals of a given species. This is a particular issue where more 'exotic' or difficult to house species are being studied, for example in zoos, or when working with threatened species in the wild (Shaw et al., 2021). Farrar and colleagues (2021) found that the median sample size of experiments in 63 animal physical cognition papers, the majority of which focussed on nonhuman primates, parrots and corvids, was just 7 individuals. However, there is also evidence that more standardised lab studies where animals are bred specifically for research are also frequently underpowered. A systematic review of around 2000 preclinical rodent experiments found that, at best, only 12.5% were sufficiently powered to detect a large effect size (Bonapersona et al., 2021). Insufficient sample sizes increase the likelihood of spurious results due to a lack of statistical power, and also may increase the likelihood that researchers will engage in QRPs to extract as many 'findings' as possible, in order to improve chances of publication (Stevens, 2017).

It is well established that unconscious observer bias can influence study results if the observer expects a particular effect (e.g, Burghardt, 2020; Burghardt et al., 2012; Forstmeier et al., 2017; Law, 2018)—a particular concern for hypothesis-testing experimental studies, but also for observational research where a specific outcome is expected. However, observer blinding appears to be relatively rare in comparative psychology and related fields. A survey that sampled articles from five animal behaviour journals (including this one) over the period of 1970-2010 found that only 6.3% included one or more instances of observer blinding (Burghardt et al., 2012). In a sample of ecology, evolution and behaviour studies published in high-impact-factor journals in 2012, data had been collected blindly in only 13% of cases (Kardish et al., 2015). This lack of blinding has been shown to have a significant impact on observations—the effect sizes of non-blind evolutionary biology studies were larger than those of blind studies (Holman et al., 2015).

Task design

Designing appropriate tasks that tap into the ability under investigation remains a central challenge today in the field of comparative cognition (Schubiger et al., 2020; Smith et al., 2018). It is hard enough to design an appropriate task for a single non-human species, never mind trying to figure out what makes a 'fair' test for diverse species, which is necessary for making valid comparisons between species—a challenge that is definitely unique to comparative psychology. Submission of articles that include direct comparisons of two or more species are explicitly encouraged by this journal, initially called for by Editor Call, and sustained by current Editor Fragaszy, and there is evidence that the publication of such papers is increasing accordingly (22% in 2007-2011, vs. 30% in 2012-2016 (Fragaszy, 2018).

As mentioned in the 'standard approach' example, developing new tasks is common in comparative cognition, yet sufficient time is rarely invested in assessing their validity—for example by assessing whether performance correlates with existing tasks purporting to measure the same ability. Thus, we rarely know for sure whether a new task actually measures the ability that we are aiming to tap into, even within a single species, never mind revalidating tasks when using them with a new species for which they were not originally designed.

Relatedly, although the use of multiple tasks within a single study can be valuable (and is also encouraged by this journal), caution is required regarding the extent to which the tasks measure the same thing. For example, Gurgand & Beran (2021) compared the performance of capuchin monkeys and human children on

'equivalent' manual and computerized detour tasks, designed to assess inhibitory control, but found no correlation in performance between the tasks for either species. Similarly, van Horik and colleagues (2018) tested pheasants reared under standardised conditions with two commonly used inhibitory control detour tasks, and found little evidence for consistent performance across the tasks, giving reason to question their construct validity.

It is well established that species-specific non-cognitive factors such as morphology, motivation, ecological relevance, and perceptual biases all have the potential to confound test performance (e.g., Kamil, 1987; Krashenninikova et al., 2020; Schubiger et al., 2020; Shaw & Schmelz, 2017; Smith et al., 2018). Some tasks are easier/harder for some species due to these contextual factors—sometimes for entirely unexpected reasons (e.g., a preference for the colour blue in North Island robins; Shaw et al., 2015).

Even within non-human primates, which are relatively closely related, subject- and task-related factors influence performance in cognitive tasks (Schubiger et al., 2020). It is not uncommon to compare much more distantly related species when addressing questions regarding the evolution of cognition (e.g., apes and corvids: Albiach-Serano et al., 2012; apes and dogs: Brauer et al., 2006), where issues of devising a fair test are amplified. Procedural differences in the testing of different species (e.g., whether or not there is a barrier between the subject and experimenter; typically, non-human primates – yes; children and dogs – no) can also be a confounding factor that has the potential to lead to invalid inferences regarding between-species differences in cognition (e.g., Clark et al., 2019; Leavens et al., 2019). Comparing the performance of nonhuman animals with human participants—which is relatively commonplace in comparative cognition in particular—raises additional issues and confounds, such as the use of verbal instruction and conspecific experimenters for humans (Smith et al., 2018).

Comparing across testing sites

When trying to compare the behaviour of different species across testing sites issues pertaining to making valid inferences regarding cross-species comparisons are exacerbated. Due to logistical constraints, only relatively few research groups have access to multiple species, which means that many of our inferences about similarities and differences between species are based on cross-site (and often cross-country) comparisons. As a result, species and test site (which incorporates many additional factors such as housing type, test history, experimenter identity, etc.) are frequently confounded.

Is there reason to be concerned about this? Unfortunately, yes: there is evidence that testing site can indeed influence a species' performance in behavioural and cognitive tasks. Even within a single strain of mice housed in tightly controlled, highly standardized lab environments, there is evidence that systematic differences in stress-like behaviours can occur across labs (Crabbe et al., 1999). A re-analysis of data from MacLean et al. (2014)—a mutli-site, multi-species comparison of performance in two inhibitory control tasks—revealed some instances of large within-species variation in performance between testing sites (Farrar et al., 2021). Squirrel monkeys' performance on the cylinder task was 60% correct in Edinburgh, compared with only 5% correct in Kyoto. The key point here is that had the squirrel monkeys in Kyoto instead been, say, capuchin monkeys, then this difference in performance would likely have been attributed to a species difference in cognitive ability, rather than a difference due to test site (Farrar et al., 2021).

Experimenter identity, which is typically confounded with testing site, has also been shown to influence rodent behaviour, even within a single lab. Individuals handled by male, but not female, experimenters showed a reduction in nociception (encoding of noxious stimulation, Sorge et al., 2014). In contrast, zebrafish behaviour appears to be unaffected by experimenter identity in a variety of paradigms (de Abreu & Kalueff, 2021). A study by Szabo et al. (2017) which de-confounded testing site and experimenter by having the same experimenter conduct a series of cognitive tasks with domestic dogs at three different sites in three different countries found that the main findings replicated across sites. Thus, while there is currently mixed evidence for the influence of testing site and associated variables (e.g., experimenter) on the behaviour of animals, clearly this needs to be given careful consideration when making inferences about differences between species.

On a different note, it should be recognised that in some areas of comparative psychology, behavioural variation between testing sites is embraced—for example by those studying cultural phenomena in wild animals (e.g., Aplin et al., 2015; Whiten et al., 1999). Documenting differences in wild chimpanzee tool-use behaviours between field sites has enabled the study of how socio-ecological conditions influence primate material culture (Koops et al., 2014).

How can we make progress? Some proposals for improvement

Having briefly described concerns regarding the reliability of research findings in psychology more broadly, and highlighted some features of comparative psychology research—and in particular experimental comparative cognition—that suggest that we as a field may have similar issues, I next discuss ways in which we can make progress, including steps we can take as individuals, research groups, and a wider community, as well as the role of institutions and gatekeepers.

Replication of existing findings

If we want to know how solid our research findings are, then presumably we should attempt to replicate key findings as other fields have done (e.g., Camerer et al., 2018; Open Science Collaboration, 2015). Although replication might be logistically feasible for accessible species that are typically housed in large numbers (e.g., rodents, fish; Yasukawa & Bonnie, 2017) replicating studies is often easier said than done. The previously described 'standard approach' in comparative cognition research does not lend itself easily to replication of studies (Beran, 2018). Few research groups focus on same question in the same species (Farrar & Ostojic, 2019), and there are logistical constraints in terms of restricted resources, such of a lack of access to required species and apparatus (Farrar et al., 2020), as well as the training and experience required to work safely with certain species (Yasukawa & Bonnie, 2017). These constraints mean that it would rarely be feasible for an independent group to replicate another group's study, even if they wanted to. Replication of field-based studies is likely to be even more challenging, especially when working with threatened species where regulatory factors may control the types of research that are permitted (Shaw et al., 2021)

Given the challenges associated with replicating comparative psychology studies, any replications are most likely to take place within-labs, and indeed the tradition in the discipline for some time was to replicate a study prior to extending it in some way to build on previous findings (Beran, 2018). In doing this researchers can ensure that they are not attempting to build on spurious results. For this reason, a study replication could be an excellent first project for a graduate student joining a research group, as well as serving as a valuable training opportunity.

It is important to bear in mind that replication should be viewed as a process of evidence-building, rather than individual replications being considered as successes or failures (Edlund et al., 2021). There are many reasons that a comparative psychology replication study might not produce the same findings, aside from the possibility that the initial findings were spurious, such as seasonal and developmental variation in behaviour (Farrar et al., 2021). In addition, the earlier section on Comparing across testing sites highlighted numerous variables that can influence animals' performance in experimental tasks, including experimenter, housing arrangements, and prior experience. Thus, even replications that do not generate the same findings as the original study are valuable as they can help researchers to identify assumptions about their methodological approach and/or target phenomenon (Halina, 2021).

On top of the logistical issues associated with conducting replications and the challenges of interpreting the findings, there is also a general lack of incentive to carry out studies of this nature, with journals typically favouring positive findings with compelling narratives (e.g., Munafo et al., 2017). Thus, on top of individual-level action, there is a need for community-level change of the structures, norms and incentives that discourage replications (see section on The role of institutions and gatekeepers).

It should be noted that even if we do conduct replication studies, Farrar and colleagues (2020) are sceptical about the potential of the field to identify false positive findings in the literature via such an approach, because the power to detect differences between original and replication studies is limited by (typically small) sample sizes in original studies. This could potentially render it infeasible for individual replication studies to enable the falsification of original claims (Farrar et al., 2020)—something that should be considered before embarking on replication attempts.

Despite the manifold challenges associated with replication studies in comparative psychology, we are starting to see encouraging signs of change. For example, in May 2021 the journal *Animal Behavior and Cognition* published a special issue (Vol 8, Issue 2) dedicated solely to replication studies and perspectives on their status and value in animal behaviour science (Brecht et al., 2021).

Design of new studies

Even for the 'standard approach' described earlier, there are steps we can take to increase the reliability of

our findings when developing new studies. For example, having many trials and using within-subjects designs provides more power to detect effects given the same resources (Farrar et al., 2020), though of course such design changes may not always be feasible, depending on the question being asked.

When developing new tasks where the intention is to compare performance in a cognitive task with another species tested previously, there is a need to balance standardisation of the core task components, while allowing for some species-specific tweaking of non-confounding factors. The extent to which changes are necessary depends on how closely related the species being compared are (Krashenninikova et al., 2020), and in most cases the aim should be to achieve functional equivalence, rather than identical methodologies (Smith et al., 2018; Tomasello & Call, 2008).

There have been some recent attempts to adapt the primate cognition test battery (PCTB, Herrmann et al., 2007)—originally developed as a standardised way to compare the cognition of great ape species—for avian species, with mixed success (Krasheninnikova et al., 2019; Pika et al., 2020). Ultimately, even when modifications are made, the test battery was originally designed to reflect the natural challenges faced by primates, meaning it lacks ecological relevance for distantly related species. Furthermore, modifications may unintentionally alter the difficulty of the task. Both of these issues may result in inappropriate inferences about the abilities of one species relative to another.

One way to address the validity issues associated with task modification is to carry out 'back testing' with modified tasks on the species they were originally designed for to ensure results are as expected and in line with existing literature (Smith et al., 2018). This may be particularly important when comparing nonhuman animals with human participants, where testing setup and procedure typically differ most between species, making it especially valuable to test humans on the adapted nonhuman animal version of a task (e.g., Brosnan et al., 2011). Although task validity is of paramount importance, incorporating these additional steps of course requires resources and individuals that may be hard to come by, and like replications studies, incentives for researchers to carry out assessments of task validity are currently lacking.

Given the evidence that unconscious observer bias can influence research findings (e.g., Holman et al., 2015; Kardish et al., 2015), when designing new studies, researchers should reflect on ways in which they might reduce these effects (Holman et al., 2015). Where possible, data should be collected blind (Burghardt, 2020; Burghardt et al., 2012). Where observer blinding is not possible—for example due to resource or logistical constraints (e.g., throughout my PhD I was the only individual conducting the studies), this should be openly acknowledged in manuscripts and the associated implications for the findings discussed.

Reporting results and making inferences

When reporting results, Stevens (2017) stresses the importance of the inclusion of effect sizes (which many journals including this one now request), and clear demarcation of confirmatory (hypothesis-testing) and exploratory (hypothesis-generating) analyses (study preregistration can facilitate this; see Adopting open science practices section below), as well as the sharing of data and code to enable others to reproduce findings for themselves (Stevens, 2017). It is also crucial that researchers avoid both under-reporting (where incomplete details of reported statistical methods/results are provided) and selective reporting ('cherry-picking', or choosing not to include results that go against the hypothesis, or are 'uninteresting'). Engaging in these practices limits readers' ability to accurately interpret the findings, as well as restricting the possibility of analytical methods being critiqued or the results reproduced (Parker et al., 2016).

Researchers should ensure that they conduct and report the results of assessments of interobserver reliability (IOR)—where an independent person who is ideally blind to the study aims/hypotheses codes a subset of the sample, and this is compared with the original coding. Burghardt et al. (2012) surveyed the articles of five animal behaviour journals between 1970-2010, including this one, and found that only 6.7% of studies mentioned IOR, and only 3.2% reported IOR assessment statistics. Given that video recording is now commonplace—both for experimental studies with captive individuals and observational studies in the field—this should be standard practice (Burghardt, 2020; Burghardt et al., 2012). One possible approach is to have undergraduate students who, for example, do not have full information about the goals of different conditions, and are not personally invested in the outcome of the project, complete the IOR coding. This can also provide a valuable behavioural coding training opportunity for students who are just starting out in the field.

We can also be more transparent in our discussion of findings, for example, by openly expressing uncertainty where necessary, to avoid drawing premature conclusions on the basis of weak evidence (Farrar et al., 2020). Making general claims about a species on the

basis of a single study/methodological approach should be avoided—for example, it may not be appropriate to extrapolate conclusions about captive individuals to their wild counterparts (and vice versa), given the known effects of captivity on behaviour and cognition (e.g., Boesch, 2021; Forss et al., 2014; Leavens et al., 2019; Tomasello & Call, 2008).

We should reflect on what evidence is required for acceptance of a claim—and in particular, extraordinary claims (Shriffin et al., 2020). For example, for particularly 'surprising' or 'exceptional' novel results (though how this would be determined given that what constitutes 'surprising' is a subjective judgment based on prior knowledge and biases is unclear), more extensive/convincing evidence might be required. More extensive evidence might consist of study replication prior to publication (see e.g., Leonard et al., 2017 for an example of this in infant cognition), or converging evidence from multiple validated measures. A tiered approach to publication that more accurately reflects the more nuanced way we think about evidence could facilitate this; for example, journals could have separate sections for the reporting of (a) results (i.e., technical reports) and (b) claims, where the results of several studies are synthesised (Shriffin et al., 2020). For comparative psychology, a section for the reporting of anecdotal and serendipitous findings which are often only discussed informally with colleagues could be valuable, given the important role these play when studying animals (e.g., Burghardt, 2013; Kamil, 1987).

Adopting open science practices

In response to the pervasive issues with the scientific research process that have cast doubt on even 'established' findings in the psychological literature, there is a growing 'open science movement', which aims to make research more transparent and reproducible (Spellman et al., 2018). This push for more open science has resulted in an explosion of new practices targeted at different stages of the research process where there is the potential for bias, ranging from study preregistration, to the sharing of reproducible code and data, to the posting of preprints to allow timely dissemination of results.

For those unfamiliar with the open science movement and its terminology, the amount of information available can be overwhelming, which can result in a state of paralysis regarding where to begin (Kathawalla et al., 2021). However, doing open science is not all-or-nothing; it is perfectly legitimate to incorporate practices over time; indeed many have advocated for a 'buffet' approach, where researchers are free to pick and choose which practices they adopt (Bergmann, 2018; Kathawalla et al., 2021). Though not specific to the field of comparative psychology, there are many accessible articles offering guidelines and practical tools for working towards open science practices that provide a useful starting point (e.g., Cruwell et al., 2019; Kathawalla et al., 2021; Nuijten, 2019).

Several researchers in comparative psychology (e.g., Beran, 2018; Farrar et al., 2020; Stevens 2017) have advocated for preregistration of experimental studies, which requires researchers to set out their methodological plans (e.g., hypotheses, study design, target sample size, data exclusion criteria, planned analyses, etc) in an online, time-stamped document prior to data collection. A link to the preregistration can be shared (anonymously) with reviewers and in the published article, so that readers are able to distinguish between what researchers planned to do prior to data collection, and what they ended up doing. Preregistration aims to reduce researcher degrees of freedom, hence decreasing the likelihood of false positive results. Even for observational studies that do not aim to test hypotheses, rendering preregistration inappropriate, there are practices that should be adopted to improve research reliability, including maintaining a reproducible workflow, stating the exploratory nature of analyses at publication, and openly sharing data and code (Ihle et al., 2017). Several journals—including this one—offer Open Science Badges to authors to acknowledge and incentivise engagement with open science practices including preregistration, as well as sharing of study data and materials. These badges are visible on published articles, signalling accessibility to readers and helping to establish new community norms (Kidwell et al., 2016).

A 'step up' from study preregistration are Registered Reports (Chambers et al., 2015); a specific format of journal article that involves the peer-review of study methods prior to data collection. This provides the dual benefit of highlighting methodological flaws before a study commences, and reducing publication bias, as the study is published regardless of its findings (neither of which is the case for standard preregistration). Despite being a relatively new article format, there is already evidence that registered reports in psychology are associated with a significantly higher proportion of reporting of null results (56%) compared with standard articles (4%; Scheel et al., 2020), with the former also being rated as more rigorous and of higher quality than

the latter when peer-reviewed by researchers (Soderberg et al., 2020). Although the number of journals offering registered reports is increasing rapidly (over 250 at the time of writing), there are currently few in the field of comparative psychology (in comparison, for example, to the field of cognitive development, which faces many similar challenges with respect to research practices). The journal *Animal Behavior and Cognition* announced the registered report format in 2018 (Vonk & Krause, 2018), but to date few researchers have engaged with this opportunity (Beran, 2020).

Some have expressed concerns that study preregistration could stifle creativity and the generation of new ideas in comparative psychology (e.g., Burghardt, 2020). It should be noted that I am not suggesting that preregistration should be a requirement of publication in our field—observational studies, exploratory work and serendipity undoubtedly have an important role to play in opening up new avenues of comparative psychology research (e.g., Kamil, 1987). Consider for example Whiten & Byrne's (1988) early work on tactical deception in primates, which comprised a series of anecdotal reports. These observations went on to be highly influential for the experimental study of animal social cognition, inspiring investigations of visual perspective taking and social problem solving in chimpanzees (e.g., Hare et al., 2000; Povinelli et al., 1990), as well as those investigating whether other species tactically deceive others (e.g., ravens, Bugnyar & Kotrschal, 2002). The aim of preregistration is by no means to exclude crucial exploratory work from the literature, but rather to make it distinguishable from hypothesis-testing studies, and thus reduce publication bias (Parker et al., 2019).

Large-scale, multi-site collaborative studies (consortium approach)

One way to address some of the aforementioned challenges associated with individual replication studies (e.g., sample size issues; access to required species; concerns regarding possible impact of testing site) is to engage in large-scale, cross-site collaboration. Collaborative projects aimed at addressing specific research questions in comparative psychology have been conducted previously (e.g., Amici et al., 2008; Amici et al., 2018; MacLean et al., 2014), but more recently there has been a move towards establishing large networks and infrastructure for ongoing collaboration. By bringing together researchers from multiple groups, sample sizes can be greatly increased, enabling more precise estimation of effect sizes, as well as the identification of factors that lead to variation across testing sites (Many Primates et al., 2019a). Such projects also aim to promote openness and transparency, and, importantly in comparative psychology, they enable questions that could not be answered by a single research group—for example, those related to the influence of phylogeny and environment on cognition—to be addressed.

This consortium approach was pioneered in other fields of psychology (social psychology: Many Labs, Klein et al., 2014; infant cognition: ManyBabies, Frank et al., 2017), but more recently infrastructure for the first large-scale, long-term collaborative project in comparative psychology—ManyPrimates—was established (Many Primates et al., 2019a; 2019b). ManyPrimates is an ongoing research collaboration between groups with access to different primate populations, which has already conducted a first study on the short-term memory of 176 individuals of 12 species housed at 11 different sites (Many Primates et al., 2019b), with further projects planned on delay of gratification and reasoning by exclusion (https://manyprimates.github.io/). To facilitate the consistency of data collection by different researchers at these sites, detailed procedural documents and videos are provided, and contributors are strongly encouraged to record a video of the setup and procedure at their site for feedback from the coordination team prior to data collection.

Researchers working on other species appear to be following suit. The ManyDogs consortium (https://manydogsproject.github.io/) will aim to replicate existing findings in the rapidly growing field of canine cognition, as well as addressing new questions related to, for example, breed and individual differences. Data collection for a first study on dogs' understanding of the human pointing gesture is underway at 13 sites, with a minimum of 16 dogs being tested at each (study preregistration: https://osf.io/gz5pj). The resulting large multi-site sample will also enable the investigation of how methodological variations (e.g., whether testing happens indoors or outdoors, the size of the room, whether or not the dog's owner is present during testing) might influence behaviour.

These large-scale, collaborative projects are undoubtedly valuable, but they are not designed to replace smaller scale research. Farrar et al. (2020) suggest, for example, that an ideal approach may be for research groups to have concurrent parallel research streams, with participation in large-scale collaborative projects being appropriate for hypothesis testing, and smaller

scale within-lab projects aimed at hypothesis generation.

The role of institutions and gatekeepers

Problems with the reliability of research across scientific disciplines are influenced by the institutions within which researchers operate—most notably journals, funding bodies and employers (Parker et al., 2016). All of these reward novelty, 'positive' results, and 'exciting' findings—via publications, grant awards, and hiring/promotion (Smaldino & McElreath, 2016; Nosek et al., 2021). Within such a system, we cannot rely solely on individual researchers to take the initiative to improve their practice; to achieve long-term change, gatekeepers of these institutions have a crucial role to play in modifying the individual-level incentives and standards for what constitutes research excellence (Schiavone et al., 2021).

There are many practical steps journals can take, such as providing checklists for researchers to adhere to in order to promote transparency and decrease bias—for example, requiring the sharing of data and analysis code (Law, 2018), and—of particular relevance to comparative psychology—mandating the reporting of IOR assessment (Burghardt et al., 2020), and requiring statements about whether or not a study was blinded (Ihle et al., 2017). Editors and reviewers can focus their evaluation of submitted papers on the validity of methods rather than the results, ensure that the claims made by authors are valid based on the data, and that speculative points are clearly demarcated (Schiavone et al., 2021).

A growing number of journals have signed up to adhere to the Transparency and Openness Promotion (TOP) guidelines developed by the Centre for Open Science (Nosek et al., 2015), which can be flexibly adopted to create an appropriate set of community standards. Some fields have gone on to adapt to TOP guidelines to facilitate transparency within field-specific constraints (e.g., the Tools for Transparency in Ecology and Evolution checklist (TEE; https://osf.io/g65cb/). Effective July 2021, this journal adopted TOP guidelines, requiring authors to include a 'Transparency and Openness' subsection in the Methods where details regarding the steps taken to comply with the guidance must be included. Authors are also encouraged to preregister their studies and include a link to this in the Author Note, together with details of the open availability of data and materials (or if they are not available, a reason should be provided).

Compared to some human psychology research that samples from the general adult population and data collection primarily takes place online and can be largely automated, data collection for both lab- and field-based comparative psychology research is relatively slow and involved. Running a study often involves extensive training of (sometimes multiple) experimenters/observers, pre-training or habituation of animals, and numerous test sessions or thousands of hours of observation. If we are to address reliability issues, for example via replication studies, then this is going to require additional time and resources, and simply would not be feasible within the existing 1—3 year timeframe of many funding awards, where there is pressure to produce novel, exciting results in order to secure publications and further funding.

Funding agencies should be concerned about the reliability of research findings and they have the potential to effect change. If awards were longer (potentially with a reduced annual budget), then there would be less pressure on researchers to quickly 'produce the goods', potentially by cutting corners and engaging in QRPs. There would be more time for replication studies, and funders could explicitly encourage these. Longer-term research programs would also enable researchers to dedicate time to many of the other reliability issues highlighted that are particularly pertinent to comparative psychology, including the assessment of task validity and the need to improve ecological validity, potentially via the combination of tightly controlled experiments with more naturalistic tasks (Smith et al., 2018), or even a combination of work with wild and captive individuals. Indeed, it has been argued that integrated research programs that incorporate a variety of settings and methodological approaches are likely to be most illuminating and impactful in comparative psychology (Snowdon & Burghardt, 2017), but short awards do not allow for the training of researchers in diverse methods that is necessary for pursuing such integrated approaches, not the time to execute them. Ultimately, slowing down science will lead not only to more reliable research, but also to a healthier, more sustainable research culture (Frith, 2020). However, institutions and gatekeepers must take the lead to make this possible.

Concluding remarks

We should all be concerned about improving the reliability of our research. Given that other areas of psychology have established that many published findings are not reliable, comparative psychology seems to be

trying to get ahead of the curve by facing the potential for replication issues in the field head-on. Although conducting direct replications is likely to be challenging in many cases due to unique features of our research, researchers are nevertheless attempting to ascertain the replicability of existing findings, and establishing infrastructure for long-term, large-scale collaborative projects. There are concrete steps we can take to improve our existing research practices, as well as some encouraging signs of structural change, but much remains to be done, particularly at an institutional level.

In sum, reflecting on our current approach to research in comparative psychology (and experimental comparative cognition in particular), acknowledging the potential for reliability issues, and being willing to take steps to improve research practice will help us in our quest to better understand the behaviour and cognition of animal species and how these may have evolved—and crucially, allow us to have confidence that our findings and the resultant claims we make are reliable.

Acknowledgements

I thank three anonymous reviewers for extremely thoughtful and constructive feedback that greatly improved the manuscript.

References

Albiach-Serrano, A., Bugnyar, T., & Call, J. (2012). Apes (*Gorilla gorilla, Pan paniscus, P. troglodytes, Pongo abelii*) versus corvids (*Corvus corax, C. corone*) in a support task: The effect of pattern and functionality. *Journal of Comparative Psychology*, *126*(4), 355-367. https://doi.org/10.1037/a0028050

Amici, F., Aureli, F., & Call, J. (2008). Fission-fusion dynamics, behavioral flexibility, and inhibitory control in primates. *Current Biology*, *18*(18), 1415-1419. https://doi.org/10.1016/j.cub.2008.08.020

Amici, F., Call, J., Watzek, J., Brosnan, S., & Aureli, F. (2018). Social inhibition and behavioural flexibility when the context changes: a comparison across six primate species. *Scientific Reports*, *8*(1), 1-9. https://doi.org/10.1038/s41598-018-21496-6

Aplin, L. M., Farine, D. R., Morand-Ferron, J., Cockburn, A., Thornton, A., & Sheldon, B. C. (2015). Experimentally induced innovations lead to persistent culture via conformity in wild birds. *Nature*, *518*(7540), 538-541. https://doi.org/10.1038/nature13998

Bachmann, C., & Kummer, H. (1980). Male assessment of female choice in hamadryas baboons. *Behavioral Ecology and Sociobiology*, *6*(4), 315-321. https://doi.org/10.1007/BF00292774

Baker, M. (2016). Reproducibility crisis. *Nature*, *533* (26), 353-366.

Beck, S. R. (2017). Interaction between comparative psychology and cognitive development. *Current Opinion in Behavioral Sciences*, *16*, 138-141. https://doi.org/10.1016/j.cobeha.2017.07.002

Beran, M. (2018) Replication and pre-registration in comparative psychology. *International Journal of Comparative Psychology*, *31*, 1-6. https://doi.org/10.46867/ijcp.2018.31.01.09

Beran, M. J. (2020). Pre-registration and assessing effects of commonly used techniques in animal behavior research. *Animal Behavior and Cognition*, *7*(*4*), 490-491. https://doi.org/10.26451/abc.07.04.01.2020

Boesch, C. (2007). What makes us human (*Homo sapiens*)? The challenge of cognitive cross-species comparison. *Journal of Comparative Psychology*, *121*(3), 227-240. https://doi.org/10.1037/0735-7036.121.3.227

Boesch, C. (2021). Identifying animal complex cognition requires natural complexity. *Iscience*, 102195. https://doi.org/10.1016/j.isci.2021.102195

Bonapersona, V., Hoijtink, H., Sarabdjitsingh, R. A., & Joëls, M. (2021). Increasing the statistical power of animal experiments with historical control data. *Nature Neuroscience*, 24, 470-477. https://doi.org/10.1038/s41593-020-00792-3

Boyle, A. (2021). Replication, uncertainty and progress in comparative cognition. *Animal Behavior and Cognition*, *8*(*2*), 296-304. https://doi.org/10.26451/abc.08.02.15.2021

Bräuer, J., Kaminski, J., Riedel, J., Call, J., & Tomasello, M. (2006). Making inferences about the location of hidden food: social dog, causal ape. *Journal of Comparative Psychology*, *120*(1), 38-47. https://doi.org/10.1037/0735-7036.120.1.38

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

Brosnan, S. F., Parrish, A., Beran, M. J., Flemming, T., Heimbauer, L., Talbot, C. F., … Wilson, B. J. (2011). Responses to the Assurance game in monkeys, apes, and humans using equivalent procedures. *Proceedings of the National Academy of Sciences*, *108*, 3442–3447. https://doi.org/10.1073/pnas.1016269108

Bugnyar, T., & Kotrschal, K. (2002). Observational learning and the raiding of food caches in ravens, *Corvus corax*: is it 'tactical' deception? *Animal Behaviour*, *64*(2), 185-195. https://doi.org/10.1006/anbe.2002.3056

Burghardt, G. M. (2013). The Janus-faced nature of comparative psychology–strength or weakness? *Evolutionary Psychology*, *11*(3), 762-780. https://doi.org/10.1177/147470491301100317

Burghardt, G. M. (2020). Insights found in century-old writings on animal behaviour and some cautions for today. *Animal Behaviour*, *164*, 241-249. https://doi.org/10.1016/j.anbehav.2020.02.010

Burghardt, G. M., Bartmess-LeVasseur, J. N., Browning, S. A., Morrison, K. E., Stec, C. L., Zachau, C. E., & Freeberg, T. M. (2012). Perspectives–minimizing observer bias in behavioral studies: a review and recommendations. *Ethology*, *118*(6), 511-517. https://doi.org/10.1111/j.1439-0310.2012.02040.x

Burley N. (1981). Sex ratio manipulation and selection for attractiveness. *Science*, *211*(4483), 721–722. https://doi.org/10.1126/science.211.4483.721

Byers-Heinlein, K., Bergmann, C., & Savalei, V. (2021). *Six solutions for more reliable infant research*. PsyArXiv. https://doi.org/10.31234/osf.io/u37fy

Caicoya, Á. L., Amici, F., Ensenyat, C., & Colell, M. (2019). Object permanence in *Giraffa camelopardalis*: First steps in giraffes' physical cognition. *Journal of Comparative Psychology*, *133*(2), 207-214. https://doi.org/10.1037/com0000142

Call, J., Burghardt, G. M., Pepperberg, I. M., Snowdon, C. T., & Zentall, T. (2017). What is comparative psychology? In J. Call, G. M. Burghardt, I. M. Pepperberg, C. T. Snowdon, & T. Zentall (Eds.), *APA handbook of comparative psychology: Basic concepts, methods, neural substrate, and behavior* (pp. 3–15). American Psychological Association. https://doi.org/10.1037/0000011-001

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., … Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*, 637-644. https://doi.org/10.1038/s41562-018-0399-z

Chambers, C. D., Dienes, Z., McIntosh, R. D., Rotshtein, P., & Willmes, K. (2015). Registered reports: realigning incentives in scientific publishing. *Cortex*, *66*, A1-A2. https://doi.org/10.1016/j.cortex.2015.03.022

Clark, H., & Leavens, D. A. (2019). Testing dogs in ape-like conditions: the effect of a barrier on dogs' performance on the object-choice task. *Animal Cognition*, *22*(6), 1063-1072. https://doi.org/10.1007/s10071-019-01297-8

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

Crüwell, S., van Doorn, J., Etz, A., Makel, M. C., Moshontz, H., Niebaum, J. C., ... & Schulte-Mecklenbeck, M. (2019). Seven easy steps to open science. *Zeitschrift für Psychologie*, , *227*(4), 237-248. https://doi.org/10.1027/2151-2604/a000387

de Abreu, M. S., & Kalueff, A. V. (2021). Of mice and zebrafish: the impact of the experimenter identity on animal behavior. *Lab Animal*, *50*(1), 7. https://doi.org/10.1038/s41684-020-00685-9

Edlund, J. E., Cuccolo, K., Irgens, M. S., Wagge, J. R., & Zlokovich, M. S. (2021). Saving science through replication studies. *Perspectives on Psychological Science*. https://doi.org/10.1177/1745691620984385

Edwards, M. A., & Roy, S. (2017). Academic research in the 21st century: Maintaining scientific integrity in a climate of perverse incentives and hypercompetition. *Environmental Engineering Science*, *34*(1), 51-61. https://doi.org/10.1089/ees.2016.0223

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

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

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

Forss, S. I., Schuppli, C., Haiden, D., Zweifel, N., & Van Schaik, C. P. (2015). Contrasting responses to novelty by wild and captive orangutans. *American Journal of Primatology*, *77*(10), 1109-1121. https://doi.org/10.1002/ajp.22445

Forstmeier, W., Wagenmakers, E. J., & Parker, T. H. (2017). Detecting and avoiding likely false-positive findings–a practical guide. *Biological Reviews*, *92*(4), 1941-1968. https://doi.org/10.1111/brv.12315

Fragaszy, D. M. (2018). Editorial. *Journal of Comparative Psychology*, 132(1), 1–3. https://doi.org/10.1037/com0000104

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., Hamlin, J. K., Hannon, E. E., Kline, M., Levelt, C., Lew-Williams, C., Nazzi, T., Panneton, R., Rabagliati, H., Soderstrom, M., Sullivan, J., Waxman, S., Yurovsky, D. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy*, 22, 421–435. https://doi.org/10.1111/infa.12182

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

Frith, U. (2020). Fast lane to slow science. *Trends in Cognitive Sciences*, *24*(1), 1-2. https://doi.org/10.1016/j.tics.2019.10.007

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "*p*-hacking" and the research hypothesis was posited ahead of time. New York, NY: Department of Statistics, Columbia University. Retrieved from http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

Giljov, A., Karenina, K., Ingram, J., & Malashichev, Y. (2017). Early expression of manual lateralization in bipedal marsupials. *Journal of Comparative Psychology*, *131*(3), 225-230. https://doi.org/10.1037/com0000073

Gurgand, L., & Beran, M. J. (2021). Assessing consistency in children's and monkeys' performance across computerized and manual detour problem tasks. *Behavioural Processes*, *182*, 104291. https://doi.org/10.1016/j.beproc.2020.104291

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

Hare, B., Call, J., Agnetta, B., & Tomasello, M. (2000). Chimpanzees know what conspecifics do and do not see. *Animal Behaviour*, *59*(4), 771-785. https://doi.org/10.1006/anbe.1999.1377

Hirsch, J. (1987). Editor's introduction. *Journal of Comparative Psychology*, *101*, 219-220. https://doi.org/10.1037/h0092595

Hodos, W., and Campbell, C. B. (1969). Scala naturae: Why there is no theory in comparative psychology. *Psychological Review*, *76*, 337-350. https://doi.org/10.1037/h0027523

Holman, L., Head, M. L., Lanfear, R., & Jennions, M. D. (2015). Evidence of experimental bias in the life sciences: why we need blind data recording. *PLoS Biology*, *13*(7), e1002190. https://doi.org/10.1371/journal.pbio.1002190

Hopper, L. M. (2017). Cognitive research in zoos. *Current Opinion in Behavioral Sciences*, *16*, 100-110. https://doi.org/10.1016/j.cobeha.2017.04.006

Ihle, M., Winney, I. S., Krystalli, A., & Croucher, M. (2017). Striving for transparent and credible research: practical guidelines for behavioral ecologists. *Behavioral Ecology*, *28*(2), 348-354. https://doi.org/10.1093/beheco/arx003

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

Kamil, A. (1987). A synthetic approach to the study of animal intelligence. *Nebraska Symposium on Motivation 1987*, 35, 257-308.

Kampis, D., Karman, P., Csibra, G., Southgate, V., & Hernik, M. (2020). *A two-lab direct replication attempt of Southgate, Senju, & Csibra (2007).* PsyArXiv. https://doi.org/10.31234/osf.io/gzy26

Kardish, M. R., Mueller, U. G., Amador-Vargas, S., Dietrich, E. I., Ma, R., Barrett, B., & Fang, C. C. (2015). Blind trust in unblinded observation in ecology, evolution, and behavior. *Frontiers in Ecology and Evolution*, *3*, 51. https://doi.org/10.3389/fevo.2015.00051

Kathawalla, U., Silverstein, P., & Syed, M. (2021). Easing into open science: A guide for graduate students and their advisors. *Collabra: Psychology*, *7*(1): 18684. https://doi.org/10.1525/collabra.18684

Kelly, C. D. (2006). Replicating empirical research in behavioral ecology: how and why it should be done but rarely ever is. *The Quarterly Review of Biology*, *81*(3), 221-236. https://doi.org/10.1086/506236

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

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196-217. https://doi.org/10.1207/s15327957pspr0203_4

Kidwell, M. C., Lazarević, L. B., Baranski, E., Hardwicke, T. E., Piechowski, S., Falkenberg, L. S., ... & Nosek, B. A. (2016). Badges to acknowledge open practices: A simple, low-cost, effective method for increasing transparency. *PLoS Biology*, *14*(5), e1002456. https://doi.org/10.1371/journal.pbio.1002456

Klein, R., Ratliff, K., Vianello, M., Adams Jr, R., Bahník, S., Bernstein, M., ... & Nosek, B. (2014). Data from investigating variation in replicability: A "many labs" replication project. *Journal of Open Psychology Data*, *2*(1), e4. https://doi.org/10.5334/jopd.ad

Koops, K., Visalberghi, E., & van Schaik, C. P. (2014). The ecology of primate material culture. *Biology Letters*, *10*(11), 20140508. https://doi.org/10.1098/rsbl.2014.0508

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

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

Kulke, L., Reiß, M., Krist, H., & Rakoczy, H. (2018a). How robust are anticipatory looking measures of Theory of Mind? Replication attempts across the life span. *Cognitive Development*, *46*, 97–111. https://doi.org/10.1016/j.cogdev.2017.09.001

Kulke, L., von Duhn, B., Schneider, D., & Rakoczy, H. (2018b). Is implicit theory of mind a real and robust phenomenon? Results from a systematic replication study. *Psychological Science*, *29*(6), 888–900. https://doi.org/10.1177/0956797617747090

Law, Y. H. (2018) Replication failures highlight biases in ecology and evolution science. *The Scientist Magazine*.

Lawrence, P. A. (2016). The last 50 years: Mismeasurement and mismanagement are impeding scientific research. *Current Topics in Developmental Biology*, *116*, 617-631. https://doi.org/10.1016/bs.ctdb.2015.12.013

Leavens, D. A., Bard, K. A., & Hopkins, W. D. (2019). The mismeasure of ape social cognition. *Animal Cognition*, *22*(4), 487-504. https://doi.org/10.1007/s10071-017-1119-1

Leonard, J. A., Lee, Y., & Schulz, L. E. (2017). Infants make more attempts to achieve a goal when they see adults persist. *Science*, *357*(6357), 1290-1294. https://doi.org/10.1126/science.aan2317

MacLean, E. L., Hare, B., Nunn, C. L., Addessi, E., Amici, F., Anderson, R. C., ... & Boogert, N. J. (2014). The evolution of self-control. *Proceedings of the National Academy of Sciences*, *111*(20), E2140-E2148. https://doi.org/10.1073/pnas.1323533111

Many Primates, Altschul, D. M., Beran, M. J., Bohn, M., Caspar, K. R., Fichtel, C., Forsterling, M., ... & Watzek, J. (2019a). Collaborative open science as a way to reproducibility and new insights in primate cognition research. *Japanese Psychological Review*, *62*(3), 205-220. https://doi.org/10.31234/osf.io/8w7zd

Many Primates, Altschul, D. M., Beran, M. J., Bohn, M., Call, J., DeTroy, S., ... & 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

Mather, J. A., & Alupay, J. S. (2016). An ethogram for Benthic Octopods (*Cephalopoda: Octopodidae*). *Journal of Comparative Psychology*, *130*(2), 109-127. https://doi.org/10.1037/com0000025

Mori, A., & Burghardt, G. M. (2017). Do tiger keelback snakes (*Rhabdophis tigrinus*) recognize how toxic they are? *Journal of Comparative Psychology*, *131*(3), 257-265. https://doi.org/10.1037/com0000075

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Du Sert, N. P., ... & Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 1-9. https://doi.org/10.1038/s41562-016-0021

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., ... & Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422-1425. https://doi.org/10.1126/science.aab2374

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Almenberg, A. D., ... & Vazire, S. (in press). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*(6), 615-631. https://doi.org/10.1177/1745691612459058

Nuijten, M. B. (2019). Practical tools and strategies for researchers to increase replicability. *Developmental Medicine & Child Neurology*, *61*(5), 535-539. https://doi.org/10.1111/dmcn.14054

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

Pallante, V., Ferrari, P. F., Gamba, M., & Palagi, E. (2019). Embracing in a female-bonded monkey species (*Theropithecus gelada*). *Journal of Comparative Psychology*, *133*(4):442-451. https://doi.org/10.1037/com0000173

Pika, S., Sima, M. J., Blum, C. R., Herrmann, E., & Mundry, R. (2020). Ravens parallel great apes in physical and social cognitive skills. *Scientific Reports*, *10*(1), 1-19. https://doi.org/10.1038/s41598-020-77060-8

Povinelli, D. J., Nelson, K. E., & Boysen, S. T. (1990). Inferences about guessing and knowing by chimpanzees (*Pan troglodytes*). *Journal of Comparative Psychology*, *104*(3), 203-210. https://doi.org/10.1037/0735-7036.104.3.203

Scheel, A. M., Schijen, M., & Lakens, D. (2021). An excess of positive results: Comparing the standard psychology literature with registered reports. *Advances in Methods and Practices in Psychological Science*. https://doi.org/10.1177/25152459211007467

Schiavone, S. R., Bottesini, J. G., & Vazire, S. (2021). *The crisis from above: Gatekeepers need better standards.* PsyArXiv. https://doi.org/10.31234/osf.io/mby5u

Schubiger, M. N., Fichtel, C., & Burkart, J. M. (2020). Validity of cognitive tests for non-human animals: Pitfalls and prospects. *Frontiers in Psychology*, *11*, 1835. https://doi.org/10.3389/fpsyg.2020.01835

Shaw, R. C., Greggor, A. L., & Plotnik, J. M. (2021). The challenges of replicating research on endangered species. *Animal Behavior and Cognition*, *8*(*2*), 240-246. https://doi.org/10.26451/abc.08.02.10.2021

Shaw, R. C., & Schmelz, M. (2017). Cognitive test batteries in animal cognition research: evaluating the past, present and future of comparative psychometrics. *Animal Cognition*, *20*(6), 1003-1018. https://doi.org/10.1007/s10071-017-1135-1

Snowdon, C. T., & Burghardt, G. M. (2017). Studying animal behavior: Integration of field and laboratory approaches. In J. Call, G. M. Burghardt, I. M. Pepperberg, C. T. Snowdon, & T. Zentall (Eds.), *APA handbook of comparative psychology: Basic concepts, methods, neural substrate, and behavior* (pp. 39–63). American Psychological Association. https://doi.org/10.1037/0000011-003

Shettleworth, S. J. (2010). *Cognition, Evolution, and Behavior* (2nd ed.). New York, NY, USA: Oxford University Press.

Shiffrin, R. M., Crystal, J. D., Wagenmakers, E. J., Chandramouli, S., Vandekerckhove, J., Zorzi, M., & Morey, R. D. (2021*). Extraordinary claims, extraordinary evidence? A discussion.* PsyArXix. https://doi.org/10.31234/osf.io/2sfbm

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. https://doi.org/10.1177/0956797611417632

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*(9), 160384. https://doi.org/10.1098/rsos.160384

Smith, M. F., Watzek, J., & Brosnan, S. F. (2018). The importance of a truly comparative methodology for comparative psychology. *International Journal of Comparative Psychology*, *31*, 37777. https://doi.org/10.46867/ijcp.2018.31.01.12

Soderberg, C. K., Errington, T. M., Schiavone, S. R., Bottesini, J. G., Singleton Thorn, F., Vazire, S., … Nosek, B. A. (2020). *Initial evidence of research quality of registered reports compared to the traditional publishing model.* MetaArXiv Preprints. https://doi.org/10.31222/osf.io/7x9vy

Sorge, R. E., Martin, L. J., Isbester, K. A., Sotocinal, S. G., Rosen, S., Tuttle, A. H., ... & Mogil, J. S. (2014). Olfactory exposure to males, including men, causes stress and related analgesia in rodents. *Nature Methods*, *11*(6), 629-632. https://doi.org/10.1038/nmeth.2935

Spellman, B. A., Gilbert, E. A., & Corker, K. S. (2018). Open science. *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience*, *5*, 1-47. https://doi.org/10.1002/9781119170174.epcn519

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

Stracke, C. M. (2020). Open science and radical solutions for diversity, equity and quality in research: A literature review of different research schools, philosophies and frameworks and their potential impact on science and education. *Radical Solutions and Open Science*, 17-37. https://doi.org/10.1007/978-981-15-4276-3_2

Szabó, D., Mills, D. S., Range, F., Virányi, Z., & Miklósi, Á. (2017). Is a local sample internationally representative? Reproducibility of four cognitive tests in family dogs across testing sites and breeds. *Animal Cognition*, *20*(6), 1019-1033. https://doi.org/10.1007/s10071-017-1133-3

Thornton, A., & Lukas, D. (2012). Individual variation in cognitive performance: Developmental and evolutionary perspectives. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1603), 2773-2783. https://doi.org/10.1098/rstb.2012.0214

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

van Horik, J. O., Langley, E. J., Whiteside, M. A., Laker, P. R., Beardsworth, C. E., & Madden, J. R. (2018). Do detour tasks provide accurate assays of inhibitory control? *Proceedings of the Royal Society B: Biological Sciences*, *285*(1875), 20180150. https://doi.org/10.1098/rspb.2018.0150

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

Vonk, J., & Shackelford, T. K. (2012). *Toward bridging gaps: Finding commonality between evolutionary and comparative psychology.* Oxford Handbook of Comparative Evolutionary Psychology, 3-16. https://doi.org/10.1093/oxfordhb/9780199738182.013.0001

Wang, D., Forstmeier, W., Ihle, M., Khadraoui, M., Jerónimo, S., Martin, K., & Kempenaers, B. (2018). Irreproducible textbook "knowledge": The effects of color bands on zebra finch fitness. *Evolution*, *72*(4), 961-976. https://doi.org/10.1111/evo.13459

Whiten, A., & Byrne, R. W. (1988). Tactical deception in primates. *Behavioral and Brain Sciences*, *11*(2), 233-244. https://doi.org/10.1017/S0140525X00049682

Whiten, A., Goodall, J., McGrew, W. C., Nishida, T., Reynolds, V., Sugiyama, Y., ... & Boesch, C. (1999). Cultures in chimpanzees. *Nature*, *399*(6737), 682-685. https://doi.org/10.1038/21415

Wynne, C. D., & Bolhuis, J. J. (2008). Minding the gap: Why there is still no theory in comparative psychology. *Behavioral and Brain Sciences*, *31*(2), 152–153. https://doi.org/10.1017/S0140525X08003786

Yasukawa, K., & Bonnie, K. E. (2017). Observational and experimental methods in comparative psychology. In J. Call, G. M. Burghardt, I. M. Pepperberg, C. T. Snowdon, & T. Zentall (Eds.), *APA handbook of comparative psychology: Basic concepts, methods, neural substrate, and behavior* (pp. 65–86). American Psychological Association. https://doi.org/10.1037/0000011-004