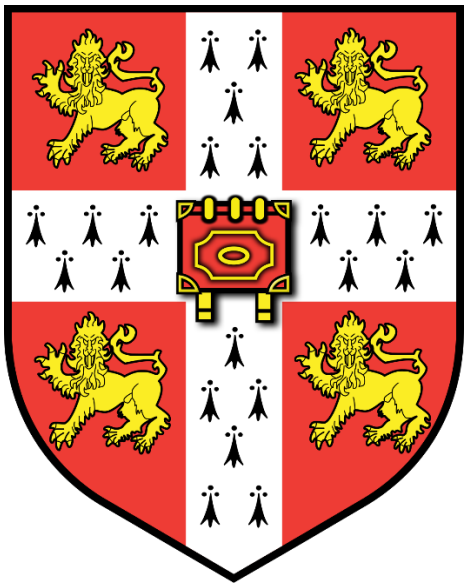
Replication, bias, and meta-research in animal cognition research



Benjamin George Farrar

Corpus Christi College

This dissertation is submitted for the degree of Doctor of Philosophy

September 2021

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text.

It does not exceed the prescribed word limit for the Faculty of Biological Sciences Degree Committee.

All chapters contain material discussed with my supervisors, Professor Nicola Clayton FRS and Dr Ljerka Ostojić. In addition:

In Chapter 2, I received formal comments on the work from Marcus Boekle and Corina Logan, and peer review from Christian Agrillo, Manuel Bohn, and an editor.

In Chapter 3, I collaborated with Konstantinos Voudouris on applying the re-sampling definition of replication to animal cognition research, and received peer review from Alfredo-Sanchez Tojar, an anonymous reviewer, and an editor.

In Chapter 4, the arguments were developed with Ljerka Ostojić.

In Chapter 5, Camille Troisi, Drew Altschul, Julia Fischer, Jolene van der Mescht, Alizée Vernouillet, Ljerka Ostojić and Sarah Placì assisted with data extraction and project discussions, and I received peer review from two anonymous reviewers and an editor.

In Chapter 6, Amalia Bastos, Alizée Vernouillet, Camille Troisi, Claudia Wascher, Elias Garcia-Pelegrin, Ed Legg, Jessie Adriaense, Katharina Brecht, Ljerka Ostojić, Mahmoud Elsherif and Sarah Placì assisted with data extraction, quality control, and project discussions. In addition, Ljerka Ostojić solicited the keywords

from social cognition experts for the literature search, performed a pilot literature search, and was the second coder for study inclusion. Mahmoud Elsherif also helped to validate the coding manual and a pilot literature search, as well as revieing 50% of the quality control data.

In Chapter 7, Alizée Vernouillet, Katharina Brecht, Elias Garcia-Pelegrin, Laurie O'Neill, Poppy Lambert, Shannon Francis, Ed Legg, Mahmoud Elsherif and Ljerka Ostojić assisted with data extraction and quality control for Study 1, as well as project discussion.

In Chapter 9, Ljerka Ostojić was the second coder for the analysis of qualitative survey responses.

Summary

Benjamin George Farrar

Replication, bias, and meta-research in animal cognition research

In this thesis I explore the extent to which researchers of animal cognition should be concerned about the reliability of its scientific results and the presence of theoretical biases across research programmes. To do so I apply and develop arguments borne in human psychology's "replication crisis" to animal cognition research and assess a range of secondary data analysis methods to detect bias across heterogeneous research programmes. After introducing these topics in Chapter 1, Chapter 2 makes the argument that areas of animal cognition research likely contain many findings that will struggle to replicate in direct replication studies. In Chapter 3, I combine two definitions of replication to outline the relationship between replication and theory testing, generalisability, representative sampling, and between-group comparisons in animal cognition. Chapter 4 then explores deeper issue in animal cognition research, examining how the academic systems that might select for research with low replicability might also select for theoretical bias across the research process. I use this argument to suggest that much of the vociferous methodological criticism in animal cognition research will be ineffective without considering how the academic incentive structure shapes animal cognition research. Chapter 5 then beings my attempt to develop methods to detect bias and critically and quantitatively synthesise evidence in animal cognition research. In Chapter 5, I led a team examining publication bias and the robustness of statistical inference in studies of animal physical cognition. Chapter 6 was a systematic review and a quantitative risk-of-bias assessment of the entire corvid social cognition literature. And in Chapter 7, I led a team assessing how researchers in animal cognition report and interpret non-significant statistical results, as well as the *p*-value distributions of non-significant results across a manually extracted dataset and an automatically extracted dataset from the animal cognition literature. Chapter 8 then reflects on the difficulties of synthesising evidence and detecting bias in animal cognition research. In Chapter 9, I present survey data of over 200 animal cognition researchers who I questioned on the topics of this thesis. Finally, Chapter 10 summarises the findings of this thesis, and discusses potential next steps for research in animal cognition.

Acknowledgements

This thesis would not have been possible without the support of many people. Most importantly, I would like to thank my parents and brother: I would be nowhere without your continued support, encouragement, and fun over the last twenty-five years.

I would like to thank my supervisors, Nicky Clayton and Ljerka Osotjić, and also Rachel Crosby. Nicky, thank-you for your unwavering support, encouragement, and kindness throughout my PhD. Ljerka, I have struggled to put into words my thanks to you. The conversations we have had over the last five years have not only shaped this thesis, but also the way I think. Your unquestionable dedication to your colleagues and students will always be an inspiration to me. Thank-you for everything. Finally, thank-you Rachel for all the discussions and craziness we had together in Madingley, I would not have made it through the PhD without you.

Thank-you to all those I have collaborated with in this thesis and beyond, it has been a pleasure to work with each of you: Alizée Vernouillet; Camille Troisi; Claudia Wascher; Elias Garcia-Pelegrin, Ed Legg; Jessie Adriaense; Katharina Brecht; Mahmoud Elsherif; Sarah Placì; Drew Altschul; Julia Fischer; Jolene van der Mescht; Laurie O'Neill; Poppy Lambert; Shannon Francis; Rachael Miller; Megan Lamber' Stephan Reber; Piero Amodio, Christopher Krupenye; Johanni Brea; Alba Motes-Rodrigo and Claudio Tennie.

Thank-you to all the peer-reviewers, editors and copy editors who worked on the published pieces of this thesis, and thank-you to the countless academics who have given up time to meet with me, host talks, and discuss the issues of this thesis. There are too many to name, but you know who you are. Thank-you also to all members and friends of the Comparative Cognition Lab in Cambridge, you have really made my PhD experience. There have been so many wonderful lab members over the years, but a special thank-you to: Corina Logan, Ed Legg, Piero Amodio, Alizée Vernouillet, Eli Garcia, Gabrielle Davidson and Rachael Miller. Thank-you also to my PhD advisors, Jacob Stegenga and Amy Orben, for their continued guidance.

Finally, thank-you to my friends for always providing fun and support over the course of the PhD. You are too many to name, but special thanks to Rob Gibbs, Ben Dunford, Orestis Sherman, Francesca Dakin, Daniel Noel, Will Moody and Sophie Hudson, who all played key roles in shaping how I think, and kept me sane throughout the PhD.

Publications

Publications included in this thesis:

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. [pre-print] *PsyArXiv.*

Farrar, B. G., Altschul, D. M., Fischer, J., van der Mescht, J., Placì, S., Troisi, C. A., ... & Ostojić, L. (2020). Trialling meta-research in comparative cognition: Claims and statistical inference in animal physical cognition. *Animal behavior and cognition*, *7*(3), 419.

Farrar B.G., Ostojić L., Clayton N.S. (2021) The hidden side of animal cognition research: Scientists' attitudes toward bias, replicability and scientific practice. PLoS ONE 16(8): e0256607

Farrar, B. G., Voudouris, K., & Clayton, N. S. (2021). Replications, Comparisons, Sampling and the Problem of Representativeness in Animal Behavior and Cognition Research. Animal Behavior and Cognition, 8 (2), 273-295

Publications not included in this thesis:

Lambert, M., Farrar, B. G., Garcia-Pelegrin, E., Reber, S. A., & Miller, R. (2021). ManyBirds: A multi-site collaborative Open Science approach to avian cognition and behaviour research. Accepted at *Animal Behavior and Cognition*

Amodio, P., Brea, J., Farrar, B. G., Ostojić, L., & Clayton, N. S. (2021). Testing two competing hypotheses for Eurasian jays' caching for the future. *Scientific Reports*, *11*(1), 1-15.

Amodio, P., Farrar, B. G., Krupenye, C., Ostojic, L., & Clayton, N. S. (2021). Little evidence that Eurasian jays protect their caches by responding to cues about a conspecific's desire and visual perspective. *Elife*, *10*, e69647.

Farrar, B. G., Krupenye, C., Motes-Rodrigo, A., Tennie, C., Fischer, J., Altschul, D., & Ostojic, L. (2021). Replication and Reproducibility in Primate Cognition Research. Book Chapter, Primate Cognitive Studies

Farrar, B. G. (2020). Evidence of tool use in a seabird? PsyArXiv, 463hk, ver. 5 recommended and peer-reviewed by Peer Community In Ecology. doi: 10.31234/osf.io/463hk

Farrar, B. G., & Ostojić, L. (2020). It's not just the animals that are STRANGE. *Learning & Behavior*, 1-2.

Farrar, B. G., & Ostojić, L. (2018). Does social distance modulate adults' egocentric biases when reasoning about false beliefs? *PloS one*, *13*(6), e0198616.

Contents

1. Chapter 1: Introduction

In 2017, when I started my PhD, psychological science was in the processes transforming its research methods. The so-called “credibility revolution” (Vazire, 2018) was catalysed by the continued failure of many landmark psychological findings to replicate (the "replication crisis" Camerer et al., 2018; Open Science Collaboration, 2015), and a growing recognition the scientific incentive structures, in some areas, actively selected for poor research (Higginson & Munafò, 2016; Smaldino & McElreath, 2016). In response to the replication crisis, a swathe of novel methods, analysis practices, editorial policies and infrastructure were being developed and implemented, largely focused on improving the robustness of scientific research. At that time, science’s replication “crisis” was well established; popular science books had been written on the topic (Chambers, 2017), and the replication crisis was even covered on HBO’s Last Week Tonight with John Oliver in May 2016 (LastWeekTonight, 2016). However, the reach of the replication crisis was not universal, with the topic being near absent from both my undergraduate and postgraduate study at the University of Cambridge until 2019, a pattern mirrored in the statistics courses of many UK institutions (TARG Meta-Research Group, 2020). The topic was also absent from the mainstream discussions in animal cognition research, other than a few key, but often subtle, voices (Beran, 2018; Craig & Abramson, 2018; Maes et al., 2016; Schubiger, 2019; Stevens, 2017; Szabó et al., 2017).

This thesis has three main parts. The first part, comprises of Chapters 2, 3 and 4, are theoretical and discuss the role of replication studies (Chapters 2 and 3), and sources of bias (Chapter 4) in animal cognition research. In the second part, Chapters 5, 6 and 7 present my attempts to develop methods capable of critical evidence synthesis in animal cognition, in physical cognition research (Chapter 5), corvid social cognition research (Chapter 6), and in interpreting non-significant findings across animal cognition research (Chapter 7). Chapter 8 reflects on the barriers I faced in Chapters 5, 6, 7, and the obstacles to critical evidence synthesis in animal cognition in general. The final part, Chapter 9, presents survey data on the attitudes of current animal cognition researchers on the topics of this thesis. Chapter then 10 summarises the thesis and discusses the many ways in which animal cognition research could develop as a science. But first, I now give a brief review of the events leading up to the “replication crisis” in human psychology and the subsequent credibility revolution it experienced. I then sketch the parallel evolution of animal cognition research up until my PhD research.

While psychology's replication crisis was not widely discussed until the early- and mid-2010s (Baker, 2016; Chambers, 2013; Gilbert et al., 2016; Nosek & Lakens, 2014; Open Science Collaboration, 2015; Pashler & Wagenmakers, 2012; Wagenmakers et al., 2011), many of the social, procedural and methodological causes of unreliable research had been discussed long before. The two most relevant to this thesis are i) the relationship between publication bias and low reliability, and ii) the distance between statistical hypotheses and theoretical claims.

1.1.1.1. Publication bias and low reliability

In 1959, Sterling reported that only 2.7% of empirical research reports contained non-significant findings across four major psychology journals in 1955. Sterling interpreted these data as suggesting that many negative results went unpublished, and thus there was a risk of the published literature simply containing many Type I errors, i.e., false positives. Sterling suggested that these false positive results could be a product of similar study designs being repeatedly used across research groups, with one eventually producing a significant result and being published. More formally, Lane and Dunlap (1978) used simulations to show that if research results are selected based on their statistical significance, then experiments overestimating the true population effect size are more likely to be published than experiments that underestimate the true population effect size. I explore this topic and emulate Lane and Dunlap's simulations for examples in comparative psychology as part of Chapter 2.

1.1.1.2. Statistical hypotheses and theoretical claims

While publication bias can lead to false positive results populating literatures, many true positive results will be published, too. However, just because researchers may correctly reject a statistical null hypothesis, this does not mean that the substantive claim that follows is justified. Meehl (1967, 1978, 1990) routinely highlighted this, noting that many tests of theory in "soft" psychological research were plagued by ten "obfuscating" factors making their results "largely uninterpretable" (Meehl, 1990). The first two of these were loose derivation chains from theory to predicted observations and problematic and untested auxiliary theories (see also Duhem, 1976). These topics are now often discussed in the context of psychology's replication crisis (Borsboom, 2014; Maatman, 2021; Oberauer & Lewandowsky, 2019; Smaldino, 2017a). Coupled with possible low statistical power to detect theoretically meaningful effect sizes, this distance between statistical hypothesis and substantive claims can mean that the likelihood of positive statistical results providing meaningful information about the construct of interest can be low.

1.2. The replication crisis

The replication crisis is most often associated with psychology, most notably within human social psychology. However, concerns about the reliability of scientific output, and often the term replication/reproducibility "crisis", have been raised across fields, from drug development (Begley & Ellis, 2012) to behaviour analysis (Locey, 2020), sports science (Halperin et al., 2018) and quantum computing (Frolov, 2021). In a survey of 1576 researchers across fields including biology, chemistry and physic, 52% of participants reported that there was a "significant" crisis in their field (Baker, 2016).

1.2.1. The proximate causes of low replicability: Bishop's four horsemen

Bishop (2019) highlighted four causes of the replication crisis, at the level of individual studies. These were publication bias (Lane & Dunlap, 1978; Sterling, 1959), low power (Button et al., 2013; Cohen, 1994), *p*-hacking (Simmons et al., 2011) and HARKing (Hypothesising After Results Are Known; Kerr, 1998). Measuring the rates of these research practices across any research programme is important to understand the evidential value of bodies of evidence, and I now introduce how researchers in disciplines other than animal cognition have attempted to measure them.

1.2.1.1. Publication bias

Publication bias is difficult to measure directly because, by definition, it is attempting to measure the rate at which studies are *not* published. This is complicated by the fact that publication bias can happen at different stages of the research process and for different reasons. For example, researchers can choose not to write up and submit their research for publication, or they might submit research for publication, but have it rejected by journals. As mentioned, Sterling (1959) "measured" publication bias by contrasting the proportion of published statistically significant results (97.3%) with the proportion of published non-significant results (2.7%). More recently, Fanelli (2010) reported that just over 90% of a sample of psychology and psychiatry papers that contained the phrase "test* the hypothes*" supported the hypothesis under investigation, and using the same method Scheel et al. (2020) found that 96% of a sample of 152 studies from Psychiatry/Psychology reported support for the hypothesis, compared to only 44% of 71 published registered reports – a publishing format in which the paper is provisionally accepted prior to data collection. Survey studies have reinforced these observational studies. For example, a survey of 454 pre-clinical animal researchers in the Netherlands found that on average they believed that 50% of studies were published (Riet et al., 2012, see Chapter 9 for animal cognition researchers' answers to a similar question).

Finally, funnel plots and *p*-value distributions can also be examined as potential indicators of publication bias (Simonsohn et al., 2014; Sterne et al., 2011). Funnel plots are often presented alongside

meta-analyses to assess the risk of publication bias and consist of a scatterplot of a measure of effect size against a measure of precision of each individual study. Asymmetric funnel plots could then be an indicator of publication bias (Sterne et al., 2011). Similarly, *p*-value distributions in which there is a drop in the number of p-values above compared to below the significance threshold might reflect a publication bias against negative results (Lakens, 2015).

Overall, converging evidence suggests that publication biases have been present across nearly all biomedical and psychological fields for much of the past 50 years. In Chapters 5, 6, 7 and 9 of this thesis, I develop and implement a range of the techniques mentioned above (proportion of positive reports, a survey study and *p*-value distributions), to provide a preliminary assessment of the degree of publication bias across animal cognition research.

1.2.1.2. Low power

The second of Bishop's four horsemen was low power research. Statistical power, in Neyman-Pearson significance testing, is the long-run probability of rejecting the null hypothesis, given a certain assumed effect size exists in the population. For example, a researcher who performs an experiment and statistical test with 75% power to detect an effect size of Cohen's $d = 0.5$ would correctly reject the null hypothesis 75% of the time in the long run, if the true effect size was 0.5.

Low power research has two downsides. The first is that researchers may make many false negative decisions, which could lead the researcher to ignore promising avenues of research (Fiedler et al., 2012), or mis-interpret the non-significance finding as evidence in favour of the null hypothesis (Aczel et al., 2018; Goodman, 2008, see Chapter 7). The second downside is that, when combined with publication bias, low powered research leads to greatly overestimated effect sizes populating the literature. This is because a low powered studies must observe very large effect sizes for the result to be statistically significant. If the study were replicated, the effect size would regress toward the "true" effect size, and this is unlikely to be significant, giving the impression of a "failed" replication (Fiedler & Prager, 2018). This is explored in Chapter 2 across various scenarios in animal cognition research.

Identifying likely low-powered research is difficult. Theoretically important effect sizes are rarely specified by researchers – and it is unclear how easy they can be derived. However, researchers have attempted to assess the likely statistical power of studies for which meta-analyses have been performed to estimate the "true" effect size. This estimation can be used to retrospectively assess the power of original studies to detect this effect size. Button et al. (2013) and Stanley et al. (2018) used this approach

to show that most individual studies included in meta-analyses across neuroscience (Button et al.) and psychology (Stanley et al.) had low power to detect the meta-analytic estimate. For example, the median power to detect the meta-analytic effect size in neuroscientific studies was 21%, and across psychological disciplines 36%. However, in both studies massive heterogeneity was observed (Nord et al., 2017; Stanley et al., 2018). Button et al. observed a "clear bimodal distribution", with the modal bin of power being 0-10% (15 meta-analyses), but seven meta-analyses also comprised studies with an average of over 90% power. As Nord et al. (2017) summarise, this heterogeneity suggests that while low power might be a large issue for many fields of neuroscientific and psychological research, it is not ubiquitous. As a microcosm of psychological research more broadly, animal cognition research is likely characterised by similar levels of heterogeneity (discussed in Chapters 2, 3, 5, 6, 7 and 8).

1.2.1.3. *p*-hacking and HARK-ing

Bishop's final two horsemen of irreproducibility were *p*-hacking and HARK-ing (Hypothesising After Results are Known; Kerr, 1998). *p*-hacking covers a range of analysis practices that might increase the chance of a false-positive result. For example, sample sizes or stopping rules and exclusion criteria may not be determined before testing begins, and then researchers may have multiple looks at the data before deciding when to stop. Researchers may record several different dependent variables, and selectively report or interpret those that 'worked' (John et al., 2012). Alternatively, they begin testing with under-specified hypotheses that could be answered through a vast number of justifiable, and often equally valid, analytical pipelines (Silberzahn et al., 2017; Steegen et al., 2016). However, only a subset of these pipelines might contain statistically significant p-values. HARK-ing similarly raises the risk of false positive results and occurs when researchers present exploratory findings as confirmatory.

Two methods have been used to quantify the rate of researchers using false positive inflating research and analysis practices: self-report surveys and meta-research or secondary data analysis projects. Self-report surveys indicate that many researchers report to have used, and suspect that others use, so-called "questionable research practices", i.e., practices that increase the frequency of false positivise (Agnoli et al., 2017; Banks et al., 2016; Fiedler & Schwarz, 2016; Fraser et al., 2018; John et al., 2012). While some of the surveys may have biased the reported results towards higher figures (see Fiedler & Schwarz, 2016), the evidence from the various surveys shows that there is a non-trivial usage of these practices. The rates of some of these practices can also be investigated through meta-research projects, an interdisciplinary approach to evaluating methods, reporting, reproducibility and the evaluation and incentives of research (Ioannidis, 2018). Thus far, meta-research and secondary data analysis projects

have largely corroborated the findings of these survey studies. To give three examples: Nieuwenhuis et al. (2011) found that 79 of 157 (50.3%) of articles interpreting interaction-like effects across the journals *Science*, *Nature*, *Nature Neuroscience*, *Neuron* and *The Journal of Neuroscience,* did so inappropriately. Gibbs and Gibbs (2015) reported articles in anaesthesiology journals frequently interpreted p-values in the range 0.05-0.10 as a trend and used to reject the null hypothesis. And in behavioural ecology, Chuard et al., (2019) found that significant *p*-values were underreported for possible confounding variables, even if the null hypothesis was true for each case (i.e., articles reported significant confounding effects at less than the false positive rate). This suggests that statistically significant confounding variables may have been selectively omitted from research reports.

Nevertheless, *p*-hacking remains difficult to detect: It is not possible to directly observe researchers performing their statistical analysis, and the literature on *p*-hacking and alpha inflation is patchy at best. Most projects focus on specific errors in specific fields (e.g., "trend" interpretation in anaesthesiology), and generalising these data to animal cognition might be difficult. In this thesis, I focus on one putative error in Chapter 7: interpreting non-significant results as evidence in favour of the null.

1.2.2. Are Most Published Research Findings False?

Ioannidis (2005, p.1) claimed "most research findings are false for most research designs and for most fields". He made this claim by calculating the probability that any given positive result was a true positive, under various assumptions about how the research was conducted (the positive predictive value, or PPV). Ioannidis' PPV was influenced by the power of a research design, alpha inflation, and the prior probability that a hypothesis was correct. For a study with 0.80 power, a pre-study odds of 50% and little bias, the probability a positive result reflects a true positive was 0.85 However, for underpowered research with higher bias and lower pre-study odds, the PPV can easily fall below 20%. And if studies are not independently replicated, without publication bias, these false positive results will likely persist in the literature (Makel et al., 2012; Nissen et al., 2016). Hence, following this argument, for a field to understand the likely informativeness of its published findings, some attempts to understand the likely statistical power, alpha inflation and plausibility of hypotheses within a field are required. Critically assessing Ioannidis' argument in relation to animal cognition research forms part of Chapter 4 of this thesis, and Chapters 5, 6, 7 and 9 provide some information on statistical power, alpha inflation and plausibility of null and alternative hypotheses in animal cognition research.

1.2.3. Evidence from large-scale replication projects

The idea that many published findings may be “false” was given credence by the results of large-scale replication projects in psychology. The most impactful of these was the Open Science Collaboration’s (OSC) ”Reproducibility Project: Psychology” (Open Science Collaboration, 2015). In this, the OSC conducted direct replication studies of 100 published social and cognitive psychology studies. Of the 97 original statistically significant findings, only 35 replications produced significant effects in the same direction as the original study. The replication studies returned, on average, half the effect size as the original reports, which forced psychologists to acknowledge that either many published effects in some areas of psychological research were not robust and/or that independently replicating true effects could be difficult.

Since then, further large scale-replication attempts have documented replicability issues across psychological research, but with large heterogeneity across research areas. In the ManyLabs series (Ebersole et al., 2016, 2020; Klein et al., 2014, 2018, 2019), direct replications were performed across multiple sites with an median sample size over 43 times larger than the original studies. Only 43 of the 77 studies (56%) produced statistically significant effects in the same direction as the original. Across all studies, the median observed effect size was 21.2% the size of the original (Nosek et al., 2021). Similarly, Camerer et al. (2018) performed direct replications with 90% power to detect 75% of the 21 experiments published in *Nature* or *Science* original study’s effect size, and, if the replication study did not produce a significant effect in the same direction as the original, they continued data collection to 90% power to detect 50% of the original effect size. Of these studies, only 13 (62%) returned significant results in the same direction as the original study, and even within these 13, the average effect size was 71% of the original studies.

1.2.3.1. Heterogeneity

While replication projects in human psychology have returned replication success rates around 60%, some projects have found areas of highly reliable research. For example, Zwaan et al. (2018) successfully replicated 9 cognitive psychology tasks (Flanker task, Simon task, Motor priming, Spacing Effect in learning, False Memory Effect, Serial Position effect, Associative priming, Repetition Priming and Shape Simulation). In the successful replications, each study produced a Bayes Factor of over 10,000 in support of the theoretical predictions over null effects. In personality psychology, Soto (2019) attempted to replicate 78 previously published personality trait-life outcome associations through four samples (all Ns in the range 1,505 – 1,559), of which 87% replicated successfully. Finally, Protzko et al. (2020) examined the prospective replicability of 16 social-behavioural science experiments using so-called “best practice”

methods: high statistical power, preregistration, methodological transparency and both internal and independent replications. They found that replication studies returned an average of 97% of the effect size as the originals, suggesting their best practice methods were sufficient to produce robust estimates.

Several factors could explain the variable replication success rate across subfields:

1.2.3.1.1. Statistical power

Subfields with higher replication rates might have performed both original studies and replication studies with higher power to detect theoretically plausible effect sizes, than fields with lower power. For example, cognitive psychology effects such as the Serial Position Effect often manifest visibly within individual participants, and so only a small sample is needed for large power. Other tasks, such as the Flanker task, use tens or hundreds of trials within individual participants, similarly increasing power (Smith & Little, 2018). In contrast, many effects in fields struggling to replicate might have been "discovered" by studies vastly under-powered to detect an underlying effect size – meaning the published effect size would be overestimated, and even the replication study underpowered to detect any underlying effect (Fiedler & Prager, 2018, but see Camerer et al. 2018)

1.2.3.1.2. Bias in original studies

If original studies in some research areas are more biased, statistically, than in other areas, their findings will be more difficult to replicate. If published findings come from a literature with a large publication bias and low powered studies, or there is frequent alpha-inflation, then replication studies will often fail.

1.2.3.1.3. Different levels of effect stability

If effects vary within and between individuals, and across time, this would place an upper limit on our expectation of the rate of replication successes. For example, in contrast to the Stroop effect, which may be consistent regardless of participants location, we might expect the effects of behavioural nudge recycling behaviour to be more context dependent, and thus more difficult to replicate (Gilbert et al., 2016).

1.2.3.1.4. Bias in replication selection

Another reason for why replication success may seem heterogenous across areas is that targets for replication are non-randomly chosen. Researchers may choose a study to replicate because its findings seem implausible, for example the claim of precognition (Bem, 2011) – a physically impossible feat – was selected for replication in a study that seemed likely to always produce negative results (Ritchie et al.,

2012). Similarly, in the large-scale replication studies candidates were non-randomly selected. For example, the studies selected by the OSC, and subsequent replication projects, were in part based off convenience (Open Science Collaboration, 2015; Klein et al., 2014, 2018). If studies are selected for replication based off doubt in the real effect, or off convenience, then it is no surprise that the replication rate will be lower than for effects with strong prior beliefs or that are less convenient to perform.

1.2.3.1.5. Bias in replication studies

Finally, replication rates might be artificially lowered due to biases or failures in the replication studies themselves. There are many ways for negative results to occur other than there being no true effect (Mitchell, 2014). For example, if an intervention delivered in person is adapted for online delivery in a new sample, this change to the intervention could reduce its salience, or even fail to elicit an effect altogether. Because there is no such thing as an identical replication study (see Chapter 3), any change to the protocol to an original study could lead to differences in study outcomes.

Chapter 2 of this thesis discusses these issues in relation to animal cognition research, asking what the field should expect from direct replication studies, if they were performed. Chapter 3 then discusses how animal cognition researchers might best conceptualise replication as it begins its own process of identifying which of its results are reliable and which are not. To set a background to this discussion, I now provide a brief comparison of animal cognition research around the 1950s in comparison to the 2000s and 2010s, and the role of replication during this (but see Beran et al., 2014 for a more thorough history).

1.3. Animal cognition research up to the replication crisis

Animal cognition research is a field that interests many but is rarely and inconsistently defined (Bayne et al., 2019). For this thesis, I take animal cognition research to be most similar to what is often labelled comparative cognition: a field that covers a wide range of topics, from how animals learn and remember to how they make decisions and how they interact with other individuals. By studying a wide number of questions in an equally wide range of species, the field broadly aims to understand the mechanisms, functions and the evolution of cognition (Beran et al., 2014; Olmstead & Kuhlmeier, 2015; Shettleworth, 2009). Related fields or definitions include comparative psychology (Call et al., 2017; Papini, 2003), cognitive ethology (Allen & Bekoff, 1997), and more distantly, behavioural ecology (Krebs & Davies, 1997) and behavioural neuroscience (Commins, 2018).

1.3.1. Animal cognition around 1950s

In the 1950s, animal cognition research was focused on studies of learning in the lab rat. In "The Snark was a Boojum", Beach (1950) lamented this limited focus of the field on the lab rat: In an analogy based on Lewis Carrol's poem "The hunting of the Snark", Beach contended that in hunting for the delicious Snark, Animal Behaviour, Comparative Psychologists instead found a deadly Boojum, the albino rat; as a result, the Comparative Psychologist faded away. Beach calculated the percentage of species and topics studied in papers between 1911 and 1948 in the *Journal of Animal Behaviour* and its successors the *Journal of Comparative Psychology*, later the *Journal of Comparative and Physiological Psychology*. Beach observed a drop in the total number of species tested, from over 20 in 1914-15 to an average of less than 10 between 1923 and 1948. Notably, the Norway Rat comprised over 50% of articles in all but one year between 1931 and 1948, and during this time the most studied topic was conditioning and learning. While Beach was critical of the state of comparative psychology at the time, he also noted the benefits of focusing on a small number of questions in a single species:

"There are many important advantages to be gained when many independent research workers attack similar problems using the same kinds of organisms... [it is] possible to check the accuracy of the findings, to accelerate the acquisition of new data, and to formulate more valid and general conclusions than could have been derived if each worker dealt with a different species." (Beach, 1950, p. 120)

Beach and others (e.g. Skinner, 1938) also noted that comparative psychology's use of the lab rat was similar to the model organism approach used in other fields, for example on Drosophila in genetics, or *Arabidopsis* in plant sciences (Krämer, 2015). Such a model organism approach embraces replication, and replication failure. For example, in 1967 Murphey claimed to demonstrate instrumental conditioning in the fruit fly, *Drosophila melanogaster*, in a maze apparatus. In this experiment, 30 male flies completed 50 acquisition and 50 reversal learning trials, in which they had to turn a specific direction in a maze apparatus to access a "reinforcement" tube extending upwards (correct response), or an aversive tube extending downwards (incorrect response). Murphey (1967) reported statistically significant effects of both acquisition learning and reversal learning. This result was exciting, because it offered a potential means to easily investigate instrumental learning mechanisms in a well-characterized genetic model. Therefore, Yeatman and Hirsch (1971) conducted a replication experiment, in which they attempted to match Murphey's protocol. They built "a replicate of the original maze", and one of the authors, Yeatman, visited Murphey's laboratory "in order to avoid gross departures from the conditions of the original study" (Yeatman & Hirsch, 1971, p. 456). Yeatman and Hirsch found no statistically significant effect of learning for the experimental group (or control groups) when following the same analysis plan as Murphey, which they interpreted as a failure to replicate the instrumental conditioning of *Drosophila*.

Similar cases of published failed replication are relatively easy to find. For example, in a 6 experiment paper, Dworkin and Miller (1986) reported a systematic failure to replicate findings of autonomic instrumental conditioning in paralysed rats. Interpreting these replication failures, Dworkin and Miller were able to discount the probability that their results were due to inadequate statistical power – something animal cognition scientists in the present often struggle to do (e.g., Amodio, Brea, et al., 2021; Amodio, Farrar, et al., 2021; Crosby, 2019). However, it would be misleading to state that inadequate replication was a non-issue in comparative psychology. In Yeatman and Hirsch's (1971) article, they discussed barriers to replication that would not have been out of place in a psychology journal today. Citing Smith (1970), they highlight "seven reasons why studies often are not replicated: (1) lack of funds, (2) lack of time, (3) lack of availability of a comparable group of subjects, (4) development of new research interests, (5) desire to publish, (6) ego involvement with the data, and (7) the reluctance of some journals to publish replication studies. " (Yeatman & Hirsch, 1971, p. 460).

1.3.2. Animal cognition in the 2000s

Since Beach (1950), and similar critiques (Bitterman, 1960; Shettleworth, 1993), animal cognition research is now broader, both in the number of species studied and the number of questions asked (Beran et al., 2014; Shettleworth, 2009). Both Beran et al. and Shettleworth, coding data from a variety of animal cognition related journals, found that rats were no longer the dominant species tested, with similar proportions of apes, monkeys, other mammals, rats and birds now being studies. More recently, a survey of the primate cognition literature (Many Primates, Altschul, Beran, Bohn, Caspar, et al., 2019) and the bird cognition literature (Lambert et al., 2021) found that between 2015 and 2020, at least 68 different primate species and 141 different bird species have been studied using cognitive tests.

This diversity of species has been matched, somewhat, by an increase in the number of topics studied. While learning studies dominated in the 1950s, today's research spans many domains, notably social cognition, physical cognition, memory and mental time travel, and personality and individual differences (Beran et al., 2014; Lambert et al., 2021; Shettleworth, 2009). While learning studies still feature, there has been somewhat of a bifurcation between animal learning and animal cognition research, with the former more often published in outlets such as *Behavioural Processes*, and the latter in journals like *Animal Cognition* and *Journal of Comparative Psychology*, amongst others. While learning research still often focuses on rats and pigeons, other cognitive studies are more distributed across taxa (Lambert et al., 2021). Animal cognition researchers have asked questions as diverse as whether frogs recognise conspecific calls (Passos et al., 2021), how otters handle stones (Bandini et al., 2021), and

whether pinyon jays and Clark's nutcrackers are differentially impacted by observers when caching (Vernouillet et al., 2021). Clearly, with such a diversity of questions to ask, and species to study them in, the probability of close, independent, replication is lower in animal cognition today than when the field had a narrower focus (even after considering the general increase in research activity – according to a Scopus search, the journals *Animal Cognition* and *Journal of Comparative Psychology* collectively published less than 100 articles each year between 1999 and 2007, but over 150 in 2018, 2019, and 2020).

Importantly, however, the diversification of animal cognition research has also increased its heterogeneity. In some areas, such as studies of hyena social cognition (Holekamp et al., 2007), cuttlefish memory (Jozet-Alves et al., 2013) or vampire bat co-operation (Carter & Wilkinson, 2013), the research is dominated by a single group of collaborators. In others, independent groups studying the same taxa converge on the same research questions, but often test different species, collaborate together, and researchers often move between the groups. This is the case, for example, in corvid social cognition research (see Chapter 7). And finally, some research areas do truly have multiple groups studying the same questions in the same species, which is the case when there is sufficient interest and sufficient opportunity to conduct this research. For example, chimpanzee theory of mind research has a high rate of conceptual independent replication (Halina, 2021) because multiple research groups have access to captive chimpanzees through collaboration with zoos, and this research area has attracted large amounts of funding. Similarly, many research groups have independently asked questions about dog social cognition (Aria et al., 2021), most likely because they are a relatively cheap convenience sample, with clear questions to ask about the effects of domestication on cognition. This unique structure of animal cognition research must be accounted for in any assessment of replication possibility, or risk-of-bias assessment, and I discuss these topics in Chapter 4, 7 and 8.

1.3.3.Top-down animal cognition and the methods wars

A feature of current animal cognition research is the prevalence of top-down research questions across many research programmes. Such top-down questions most often start with a cognitive ability and then seek to test or demonstrate this ability in an individual or group. Such research often uses null hypothesis significance testing, and leads to a dichotomous claim about some feature of animal cognition, such as "Ravens attribute visual access to unseen competitors" (Bugnyar et al., 2016), "Great apes anticipate that others will act according to false beliefs" (Krupenye et al., 2016), and "Cuttlefish show flexible and future-dependent foraging cognition" (Billard et al., 2020). Such claims depend on the reliability of the statistical result, and the validity of the operationalisation, or closeness between the

substantive claim and the statistical hypothesis. Animal cognition researchers expend great energy in debating the suitability of specific tasks to test certain cognitive abilities, both formally through commentaries and discussion in papers (e.g., Amodio et al., 2019; Beran, 2015; Farrar, 2020; Heyes, 2017; Vonk, 2018, 2019a, 2019b, 2019b), but also in the uncounted hours spent peer-reviewing papers, across multiple journals. And less frequently, researchers write articles critiquing the approaches of entire research programmes (e.g., Anderson & Gallup, 2015; Heyes, 2012, 2015; Leavens et al., 2019; Lind, 2018; Smith et al., 2014), of specific issues within the field (e.g., Andrews & Huss, 2014; Barker & Povinelli, 2019; Beran, 2012; Buckner, 2013; Burghardt et al., 2012), or of the field's approach as a whole (e.g., Allen, 2014; Eaton et al., 2018; Vonk, 2021).

In animal cognition research, the same data are often interpreted in contrasting ways by scientists (Boyle, 2021). When authors make dichotomous claims about the presence or absence of cognitive abilities in animals, others will often disagree. Occasionally, researchers label others' interpretations as unduly biased, either in favour of exceptional cognitive abilities in animals (e.g., Hill, 2017; Penn & Povinelli, 2013) or equally as being overskeptical (e.g., Fitzpatrick, 2008). In Chapter 4 of this thesis, I form my own argument about why there is so much methodological criticism in animal cognition research, but little agreed up on progress. I then detail how bias can direct research programmes, but instead of focusing on the specifics of interpreting the results of individual experiments, I focus on explaining how the academic incentive structure determines how animal cognition research is performed, and why I believe this is sufficient to explain why, i) animal cognition research faces a replication crisis, but also, ii) why many areas of animal cognition research struggle to make discoveries that convince the "killjoys" and skeptics.

1.4. Overview of thesis

In this thesis I aim to explore how concerned animal cognition researcher should be about the reliability of its scientific results and the presence of theoretical biases across research programmes. In **Chapter 2**, I start by asking what animal cognition research should expect from direct replication studies and suggest that areas of animal cognition research likely contain many findings that will struggle to replicate. In **Chapter 3**, I then combine two definitions of replication to outline the relationship between replication and theory testing, generalisability, representative sampling, and between-group comparisons in animal cognition. **Chapter 4** explores deeper issue in animal cognition research, examining how the academic systems that might select for research with low replicability might also select for theoretical bias across the research process. Following this, I focus on one key general issue facing animal cognition

researchers: critical and quantitative evidence synthesis and risk of bias assessments. In **Chapter 5**, I led a team examining publication bias and the robustness of statistical inference in studies of animal physical cognition. **Chapter 6** provides a systematic review and a quantitative risk-of-bias assessment of corvid social cognition literature. And in **Chapter 7**, I assess how researchers in animal cognition report and interpret non-significant statistical results, as well as the *p*-value distributions of non-significant results across a manually extracted dataset and an automatically extracted dataset from the animal cognition literature. **Chapter 8** then discusses the difficulty animal cognition researchers face when critically synthesising evidence and detecting bias, and I highlight this as the major challenge for animal cognition research alongside assessing the reliability and validity of individual findings. In **Chapter 9**, I present survey data of over 200 animal cognition researchers who I questioned on the topics of this thesis. Finally, **Chapter 10** summarises the findings of this thesis, and discusses potential next steps for research in animal cognition.

2. Chapter 2: A replication crisis in animal cognition?[1]

In this chapter, I question why animal cognition research has not yet experienced a replication crisis and ask what the field might expect if many of its studies were replicated. The chapter focuses on "direct" replication studies – studies where an original protocol is followed as closely as possible (D. J. Simons, 2014), but exactly what a replication study is, and how this relates to theory testing, is discussed in Chapter 3.

2.1. Why has animal cognition not yet experienced a replication crisis?

The replication crisis of human psychology had many antecedents. Most often cited are the cases of Daryl Bem and Diedrick Stapel, the publications of Ioannidis (2005) and Simmons et al., (2011), and the Open Science Collaboration's Reproducibility Project: Psychology (2015). While Ioannidis and Simmons et al., and the associated discussion, is directly translatable to animal cognition research, Bem and Stapel are cases with no clear homologues in animal cognition. Bem (2011) published a series of nine experiments in the American Psychological Association's *Journal of Social and Personality Psychology.* These nine experiments provided "evidence" in favour of extra-sensory perception, namely precognition — that a future event could affect an individual's cognition before it had occurred. While this is physically impossible, it was published in a leading journal as "Experimental evidence for anomalous retroactive influences on cognition and affect". However, Bem had followed all required protocols for making scientific claims at the time; he had conducted experiments, performed statistics, and sent these off for peer review. The issue for psychologists is that, if evidence for the clearly false precognition can be published in this system, how many other incorrect, but more plausible findings, have made it into the literature? However, Bem's case did not just highlight issues with evidential standards in psychology, but also with the approach and intentions of researchers in the field, as quotes from Bem himself illustrate:

"If you looked at all my past experiments, they were always rhetorical devices. I gathered data to show how my point would be made. I used data as a point of persuasion, and I never really worried about, 'Will this replicate or will this not?'" (Bem, quoted in an interview with Engber, 2017)

[1] This chapter contains material published in 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), 1-2 and parts of Farrar, B. G., Krupenye, C., Motes-Rodrigo, A., Tennie, C., Fischer, J., Altschul, D., & Ostojic, L. (2021). Replication and Reproducibility in Primate Cognition Research. Book Chapter, Primate Cognitive Studies

"Examine [the data] from every angle. Analyze the sexes separately. Make up new composite indexes. If a datum suggests a new hypothesis, try to find additional evidence for it elsewhere in the data. If you see dim traces of interesting patterns, try to reorganize the data to bring them into bolder relief. If there are participants you don't like, or trials, observers, or interviewers who gave you anomalous results, drop them (temporarily). Go on a fishing expedition for something—anything —interesting. No, this is not immoral. The rules of scientific and statistical inference that we overlearn in graduate school apply to the 'Context of Justification'." (Bem, 2004, p. 2)

Bem's case pointed to deeper incentive issues in the scientific process: if experiments and data analysis were mere rhetorical devices for researchers to make big claims, then why should we trust the published literature? That researchers might be incentivised to chase claims over evidence was then exemplified by the case of Deidrick Stapel, the Dutch researcher who admitted to the long-term fabrication of data across dozens of social psychology experiments (Stapel, 2014). The pressure to publish in high-impact outlets features prominently in Stapel's confessions:

"Get writing, get publishing, otherwise you know where the door is... Everyone knows that they have to publish, and everyone knows the hierarchy of the journals that they're aiming for." (Stapel, 2014, p. 87)

Accusations of misconduct have occurred in animal cognition research. For example, a committee of the Harvard Faculty of Art and Sciences found Marc Hauser "solely responsible" for eight instances of research misconduct (*Harvard Magazine, 2010)*, however the details of this misconduct are limited. The retraction notice of a paper implicated by this misconduct simply states that:

"This article has been retracted at the request of the authors. An internal investigation at Harvard University of the reported research found that the data do not support the reported findings. The authors are therefore retracting this article. M. Hauser accepts responsibility for the error." (retraction notice to Hauser et al., 2002, p.1)

More recently, data irregularities, including possible duplications, were noticed in the work of Jonathan Pruitt, a researcher of spider personality (Bolnick, 2021). While Pruitt denied misconduct (Pennisi, 2020), the behavioural ecology community were sceptical that the Pruitt data could have arisen by mistakes alone. For example, Niels Dingemanse commented on the *Science* article covering the Pruitt story, stating that: "I and my colleagues find it hard to reconcile Pruitt's data irregularities as mistakes." As of February 2020, seventeen of Pruitt's papers were either i) in the process of being retracted, (7) ii)

requested to be retracted by co-authors (5), or iii) flagged as containing possible data anomalies (5) (Viglione, 2020).

2.1.1. Why large-scale replication projects are unlikely in animal cognition

However, the cases of Hauser and Pruitt did not spark the animal cognition research community into performing systematic replication projects like those described in human sciences – the projects that unequivocally confirmed the "replication crisis", and perhaps for good reason. Three factors make such a study unlikely in animal cognition. First, researchers are unlikely to have access to the species needed to perform a replication study. For example, replication projects of cognitive studies in captive Eurasian jays are only being performed at the University of Cambridge by trained researchers affiliated with the University (Amodio, 2020; Crosby, 2019), because this is the only place to house these animals. Second, ethical approval may prove a barrier for some replication studies in animal cognition, as there is likely more heterogeneity between ethical approval requirements for animal research between institutions and nations, than for human research. Finally, animal experiments are costly, from animal housing and husbandry costs to the human costs of often long-term habituation and training protocols. This is in great contrast to the large-scale replication projects in human psychology, in which replication targets are often selected partly based on the ease of adapting a protocol for online or battery-testing (Klein et al., 2014, 2018; Open Science Collaboration, 2015)

2.1.2. The rate of replication in animal cognition research

While large-scale or systematic replication projects have not occurred in animal cognition research, this does not mean that replications never occur. According to data compiled by the ManyPrimates project (2019), 8.7% (50/574) of primate cognition studies published from January 2014 to October 2019 were replication studies, which they defined as studies that tested different populations of the same species with the same methodology (i.e., direct replications). Notably, less than one percent (0.6%, 4/574) of studies were within-paper replications, in which the authors conducted and reported replication studies within an individual publication (for examples, see Forss et al., 2019; Krupenye et al., 2016; Wallace et al., 2017). Similarly, the ManyBirds collaboration found that 42 out of 567 (7.4%) surveyed bird cognition studies from 2015 to 2020 were self-defined as replication attempts by the authors (Lambert et al., 2021). These data suggest that direct replication is not a routine aspect of primate or avian cognition research. However, the rate of replication likely differs between study designs, laboratories and individual researchers. Due to this heterogeneity, it is difficult to interpret what these "rates" of replication figures mean, other than that clearly direct replication studies likely make up less

than 10% of the published literature. However, this does not mean that replications are performed uniformly across the literature—it is likely that some research programmes very rarely publish replication studies, and some publish them at a rate much greater than 10%.

When interpreting replication rates, we must also ask what type of study designs are being replicated, and why. Logically, tasks that are low cost, quick and adaptable may be more likely to be replicated across time and sites than those that are expensive, slow and difficult to adapt. Accordingly, tasks using few and simple apparatuses with little training requirements appear to be those that are replicated most often. For example, tasks using simple tube apparatuses, such as trap-tube tasks and tube tasks for handedness, make up a significant proportion of replication studies in primates. In the tube task, which has been replicated several times both within and between species (Chapelain et al., 2011; Hopkins et al., 2004; Llorente et al., 2011; Motes Rodrigo et al., 2018; Nelson et al., 2015), a tube with two openings and a food reward smeared in the middle is provided to the test subjects in order to assess which hand they use to retrieve the food. As such, the task can be applied easily and repeatedly to nearly all testable primate groups.

A positive feedback loop may then exist whereby tasks that are used in replication studies are more likely to be subject to further replication attempts. As more and more data become available on certain tasks, interpretative frameworks can be built around them and researchers can more easily produce a narrative around their data; more comparisons are possible, the results are more easily contextualized, and less work may be needed to justify the task design (Latour & Woolgar, 1986). This feedback loop may partly explain the frequency at which test-batteries (Herrmann et al., 2010; Schmitt et al., 2012), mirror-response studies (Anderson & Gallup, 2011), inhibition tests (MacLean et al., 2014), tests of spatial memory with arrays of cups (Many Primates, Altschul, Beran, Bohn, Call, et al., 2019), and quantity judgement tests using food sets (Beran, 2001) are replicated. While biases towards replicating simpler and more popular tasks may exist, this is not necessarily detrimental for scientific progress. In fact, focusing replication attempts on tasks that are easy to perform can be justified from both the perspective of productivity and informativeness (see Krasheninnikova et al., 2020, for the case of test batteries). Simpler tasks allow researchers to collect more data from more samples of animals, which then allow for more comparisons to be made (Beach, 1950). Simpler tasks can also allow for multisite data that can probe the representative of heterogeneous populations to be collected (Many Primates, Altschul, Beran, Bohn, Call, et al., 2019).

2.1.3. What should animal cognition expect from direct replication studies?

In light of the replication crisis in human psychology and the unknown rate of replication (and replciation success) in animal cognition research, this chapter now focuses on two questions. First, what should animal cognition research expect from direct replication studies, like those originally perfomed at the onset of the human replication crisis? And second, how can we use this information to understand and develop animal cognition as a science? To address these questions, I apply arguments that have been made across fields in response to their replication crises or "credibility revolutions" (Gelman & Geurts, 2017; Vazire, 2018). While the paper focuses on what animal cognition should expect from replication studies, and how it can use this knowledge to improve in the future, the arguments can also be applied retrospectively to assess the evidential value of published findings. This is important as the established scientific system can select for unreliable and misleading research (Higginson & Munafò, 2016; Lazebnik, 2018; Lilienfeld, 2017; McElreath & Smaldino, 2015; see Chapter 4).

First, I examine statistical cues to replicability, and specifically ask what can be gained from looking at *p*-values and effect sizes in published findings. Next I ask whether the different research designs employed across animal cognition may lead to systematic differences in replicability. Then, I make the case that single direct replication studies will be unable to falsify many claims in animal cognition, even if they are false, and that because the probability of independent replicaiton is so low for much animal cognition research, most findings are practically unfalsifiable. Next I discuss the difficulties that animal cognition will face when performing and interpreting the results of replication studies. Across this discussion, I highlight the 10 conclusions I made during the beginning of my PhD which provide a platform for the deeper disucssions of replication (Chapter 3), bias (Chapter 4) and evidence synthesis (Chapters 5, 6, 7 and 8) that occurred later in my research.

2.1.3.1. Statistical cues of replicability: *p*-values, effect sizes and direct replications

Some cues about a result's likely replicability can be found in statistical reports, in the reported *p*-values, effect sizes and confidence intervals. However, the statistics presented in published reports are unlikely to provide an accurate estimate of an effect size because of a publication bias favouring positive results. If papers are selected for publication based on a significance criterion such as $p < 0.05$, then it is more likely that studies overestimating effect sizes will be published than studies that underestimate the effects (Cumming, 2008; Fiedler & Prager, 2018; Hedges, 1984; D. M. Lane & Dunlap, 1978; Piper et al., 2019; Vasishth et al., 2018). In cases where many low powered studies are performed, as we might expect in some animal cognition research, effect sizes can be overestimated upwards of 100%, and the likelihood

of an exact replication study producing a significant effect in the same direction as the original study can be low, as my later simulation studies will show. But first, to illustrate this descriptively, I now discuss three hypothetical scenarios of a direct replication study, in which the original study estimated the unknown true effect size either exactly, underestimated it or overestimated it. These direct replication studies have the same sample size as the original study and sample from the same population.

Scenario 1 – original study estimated the real effect size exactly

If an original study estimated the real effect size for that design exactly, then the likelihood of the replication study to return a positive result is the power of the statistical test to detect this effect size. Therefore, if the original study returned a positive result that was just statistically significant, such as p = .049, then the likelihood of an exact replication study returning a positive result is approximately 50%; like tossing a coin (Piper et al., 2019). This is because, 50% of the time the replication study will overestimate the effect size, resulting in $p < .049$, whereas 50% of the time it will underestimate the effect size, resulting in $p > .049$, and this is non-significant with an alpha of .05. However, as the p-value of the original study decreases, then *ceteris paribus*, the chances of a positive replication result increase.

Scenario 2 – original study underestimated the real effect size

If the original study underestimated the real effect size, then it becomes more likely that a direct replication study of equal sample size will return a positive result as the replication study will exceed that effect size over 50% of the time. In the example of a just significant result, p = .049, in the original study, the likelihood of a positive direct replication with the same sample size is now greater than 50%.

Scenario 3 – original study overestimated the real effect size

However, the most likely scenario is that the original study will have overestimated the real effect size, particularly for studies that produced p-values just below the alpha level. This scenario is most likely for two reasons. First, as publications are selected based on a cut-off point, studies that overestimate effect sizes are more likely to be published than studies which underestimate effect sizes (Hedges, 1984). For example, if exactly estimating the "true" effect size would have yielded a p-value of .06 in a design, then only samples overestimating the effect size would be published. Hence, in direct replication studies, the replication p-value would regress to the mean, .06, and the original study would have less than a 50% chance of a positive direct replication with the same sample size. Second, the overestimation of published effect sizes is further exacerbated if research practices that produce more significant results are used, which appear relatively common (Agnoli et al., 2017; Fiedler & Schwarz, 2016; Fraser et al., 2018; Simmons et al., 2011; see Chapters 4 an 9). Hence, the default expectation when directly replicating a study with a

p-value in the range of .01 < *p* < .05 is that there is a good possibility that a direct replication study with the same sample size will not produce a statistically significant result in the same direction as the original study, even if there is a real underlying effect.

2.1.3.1.1. Simulation study

Another way to view these ideas is through simulation studies, and to this end I simulated a very simple model of animal cognition research.

2.1.3.1.1.1. Methods

First, I simulated 40,000 studies, comprising of four sets of 10,000 studies with a power of 80%, 50%, 20% and 5% (i.e., false positive results only in the 5% group) to detect a known difference between two groups of 10 animals. The groups were compared using a one-tailed two-sample t-test and results with $p < .05$ were termed significant and "published."

The code for the simulations is available at: https://github.com/BGFarrar/P-value-simulations/blob/master/CCreplicationsV1.R). Data were simulated from two normal distributions for each of the four sets of simulations (Table 1).

Table 1: Details of the simulation populations.

Power	Population 1	Population 2
80%	$X \sim N(50, 5)$	$X \sim N(55.78, 5)$
50%	$X \sim N(50, 5)$	$X \sim N(53.82, 5)$
20%	$X \sim N(50, 5)$	$X \sim N(51.87, 5)$
5%	$X \sim N(50, 5)$	$X \sim N(50, 5)$

Note, details of the distributions are in the form $X \sim N(\mu, \sigma)$, where μ = mean and σ = standard deviation of a given behaviour measured in a population.

The difference between Population 1 and Population 2 was calculated in order to give the desired power for a one-tailed two sample t-test with n = 10 per group. 10,000 samples were then taken from each population and compared to each other, and the *p*-values and mean difference between each sample recorded. The proportion of *p*-values under 0.05 was calculated, and the mean difference between samples associated with these *p*-values was compared to the mean difference across all samples to calculate the unstandarised effect size inflation. Next, the expected number of exact replication studies that produced a significant result in the same direction as the original was calculated by multiplying the

number of significant results from the simulation by the power of test again, and this was performed for a range of *p*-values, as well as overall.

2.1.3.1.1.2. Restults

The results of these hypothetical studies are displayed in Table *2*, along with a comparison of the specified and "published" effect sizes.

Table 2: The results of 40,000 simulated animal cognition studies by power. The proportion published represents the proportion of studies producing $p < .05$, and the unstandardised effect sizes are the mean differences between the groups.

Power	Proportion Published	Unstandardised Effect Size		
		All Samples	Published	Mean effect size overestimate in "published" findings
80	0.796	5.77	6.53	13%
50	0.494	3.82	5.56	45%
20	0.205	1.89	4.88	158%
5	0.053	0.007	4.45	64486%

Note, the differences between the all samples effect size and the true population values specified in Table 1 occur because of sampling variance in the simution, which may be accentuated by rounding.

In total, 15,471 of the 40,000 simulated studies produced significant results, with approximately half of these from the 80% powered designs (Table *2*). However, even when running studies with 80% power, the average effect size was inflated as only the significant studies were published. As the power of the tests decreased, the degree of inflation increased. Next, I derived the expected proportion of direct replication studies that would return a significant result if all of the 15,471 published studies from the simulation were replicated exactly. Again, these exact replications had equal sample size to the original studies, and the results are broken down across different ranges of the published *p*-values in Table *3*. The overall expected replication rate of this group of research, conducted with a quarter of studies having 80%, 50%, 20% and 5% power was 0.60.

Table 3: The mathematically derived probability of a successful replication attempt of an original study randomly selected from a given range of p-values from the 15,471 "published" simulation studies.

P-value original study	$p \leq 0.01$	$0.01 < p \leq 0.02$	$0.02 < p \leq 0.03$	$0.03 < p \leq 0.04$	$0.04 < p \leq 0.05$
Probability of successful replication	0.67	0.57	0.52	0.50	0.48

These results show that the expected replication rate of published research from this model can be close to 0.5, or even below 0.5, for just significant results. Notably, these results are in the absence of any false positive inflating research practices (Fraser et al., 2018; John et al., 2012; Simmons et al., 2011); they are solely the consequence of only publishing research with $p < 0.05$ and performing at least some research with relatively low power. However, this simulation does not accurately characterize the field of animal cognition: Not all animal cognition research is performed with 80%, 50%, 20% or 5% power, or using two-sample *t*-tests comparing two groups of 10 animals. As such the numbers should not be taken as a literal estimate of the replicability of animal cognition research; rather the simulation study shows that the research and publication methods used in animal cognition do lead to effect sizes being over-estimated and a replication success likely to be closer to 50% than 100% for just significant published findings. Conversely, experiments reporting very low *p*-values, with associated effect sizes and confidence intervals that are narrow and far from 0 can be good indicators of a reliable statistical effect[2]. Finally, one scenario in which these conclusions might be inappropriate is when considering research using designs and test combinations in which the *p*-value distribution is not uniform or near uniform under the null hypothesis, such as with a binomial test and a small number of observations (see Chapters 5, 6 and 7).

[2] That is, under conditions of publication bias, many just-significant *p*-values across a research area is a good indicator of bias. If there are no biasing selection effects, however, p-values from bodies of research performing Neyman-Pearson null-hypothesis signficance testing should be intetrepted relative to the pre-specified long-run error rates, alpha and beta, with beta being calculated in relation to the minimun theoreitcally important effect size.

Conclusion 1: Animal cognition research should not expect just significant initial findings to replicate consistently, and published effect sizes are likely to be overestimated because of a publication bias towards positive results

The following section now focuses on whether the differences in research methods across animal cognition, that were not considered in the simulation study, can allow us to estimate the replicability of research in animal cognition.

2.1.3.2. Replications across animal cognition: What can we expect?

The simulation study showed that, all else being equal, studies in animal cognition with lower *p*-values are more likely to replicate than studies with higher *p*-values. However, this begs the question: which features of animal cognition research lead to effect sizes large enough to be reliably detected, and thus produce low *p*-values and successful replication studies?

One of the areas in which animal cognition research methods differ markedly is in the number of trials they use, for example experiments using touchscreens and short time periods can use hundreds or thousands of trials (e.g., Beran, 2006) whereas experiments using more constrained behaviors such as food caching usually use considerably fewer trials (Ostojić et al., 2017). Given that sample sizes are often small in animal cognition, many "effects" might be too small to detect on a single trial, as the errors associated with measuring these behaviors on a single trial are large. Having many trials in within-subjects designs can address this issue by enabling more precise estimation and consequently smaller *p*-values and more replicable results to manifest. All things being equal, within-subject designs are more powerful to detect effects than between-subject designs with similar resources, and studies with more trials will be more likely to detect effects than those with few trials. This tendency of within-subject design with many trials to be more replicable is visible in the human replication literature (Open Science Collaboration, 2015; Zwaan, Pecher, et al., 2018). Many reliable findings in animal cognition will have harnessed this ability to perform repeated tests on individual animals that are often highly trained on the task, although whether a certain number of trials are "sufficient" or not will depend on particular features of the task at hand, notably the size of the effect and the measurement error.

Conclusion 2: Studies using within-subject designs with many trials and less-noisy behaviors are more likely to produce replicable results than studies with similar resources that use between-subjects designs, within-subject designs with few trials, or more-noisy behaviors.

To illustrate this point in an applied setting, I performed another simulation based on a real experiment. Emery and Bird (2009) reported that seven rooks looked longer at two images of physically impossible events than two images of physically possible events. On average, the rooks looked for around 1000ms at each image, and approximately 200ms longer at the physically impossible events than the physically possible ones (data from Emery and Bird, 2009: Experiment 1).

2.1.3.2.1. Methods

To investigate the role of trial number and effect size on the replicability of significant results in experiments like this, I simulated data for a series of hypothetical experiments, in which seven rooks were given either 1, 5 or 100 trials in each condition, and where the average effect size was either 200ms, 100ms or 50ms. I decided that 200ms is likely quite a large effect, as it was detected in a sample size of 7 from two trials per condition, and so also included two smaller effects in the simulation. For each design I ran 10,000 simulations to estimate the power of the design and test combination. In the data simulation, I included a fixed main effect of condition, random intercepts and slopes for each individual, and a small random intercept for the stimuli. Again the code for the second simulation can be found at: https://github.com/BGFarrar/P-value-simulations/blob/master/CCreplicationsV1.R

The data were simulated using edited code from DeBruine and Barr (2019). Data were simulated from the following model:

$$LT_{si} = \beta_0 + S_{0s} + I_{0i} + (\beta_1 + S_{1s})X_i + e_{si}$$

This model is identical to DeBruine & Barr (2019, p. 7), but I swapped LT (Looking Time) for RT (Reaction Time). The looking time for subject s on item i, LT_{si}, is composed of a population grand mean β_0, a by-subject random intercept S_{0s}, a by-item (either physically possible or impossible image) random intercept I_{0i}, a fixed slope β_1, a by-subject random slope S_{1s}, and a trial-level residual e_{si}. X_i is the condition.

Across all simulations, the following parameters were simulated with the following:

β_0: 1000

S_{0s} ~ *N*(0, 100)

I_{0i} ~ *N*(0, 5)

S_{1s} ~ *N*(0, 40)

e_{si}: 200

Subjects were simulated with a correlation between intercepts and slopes of 0.2, meaning that subjects with larger looking times showed on average larger looking time differences. Across the simulations I varied the main effect of condition and the number of trials in a 3 x 3 design:

β_1 (200, 100, 0) x trials (1, 5, 100)

10,000 datasets were simulated for each design, and the analyses differed slightly between the designs to avoid singular fits. For the single trial designs, the data were analysed using paired *t*-tests, and for the five and one hundred trial designs, the data were analysed using a mixed effect model with the following structure:

lmer(LookingTime ~ Condition + (1 | subj_id), simulateddata, REML = FALSE)

Finally, for the five trial conditions, the calculated *p*-values might be slightly inaccurate as a small proportion of simulations still led to singular fits. This may also be something to consider when interpreting the analysis of Bird and Emery (Bird & Emery, 2010), which has even more parameters in the model.

2.1.3.2.2. Results

Figure *1* plots an example of the data from one of each of the designs, and the power for each design is presented under each graph. Figure 1 shows that both increasing trial number and studying larger effects can lead to large increases in power, even with a constant sample size. When there was only 1 trial per condition (leftmost panels), the power of the design to detect the whole range of effect sizes was low. Notably, this design was prone to producing results *in the opposite direction* to the true effect (see Gelman & Carlin, 2014). For the smallest effect size, 50ms, 32% of single trial experiments led to effects estimated in the opposite direction to the true effect, and 13% of the significant results were in the opposite direction to the true effect. This design was completely insensitive to individual differences in the effect, and the significant results from these studies, if published, would lead to large overestimations of the true effect size and exact replication studies often would return non-Signiant results, at the rate of 1 – power (as per the results of the earlier simulation). One striking aspect of Figure 1 is just how misleading visualisations of single trial studies can be. As the uncertainty is never measured for each bird in each condition, it cannot be visualised.

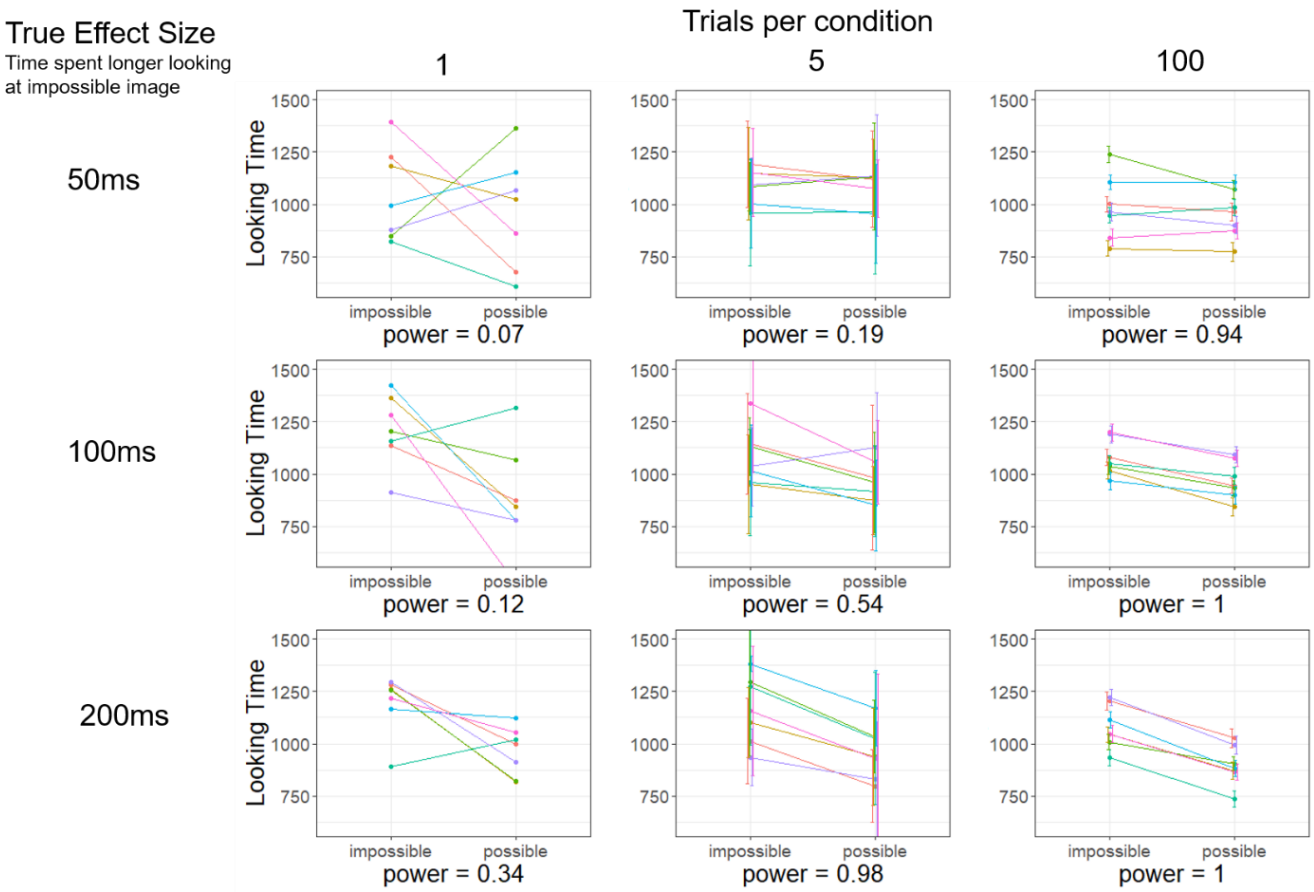


Figure 1: Graphs showing example data from nine different designs of a looking-time experiment in rooks. These designs vary on the underlying effect size (50ms, 100ms, 200ms), and the number of trials per condition (1, 5, 100). N=7 for all designs. The power of each design is printed below each graph and was calculated by simulating 10,000 studies in each design and calculating the proportion of p-values less than 0.05.

The benefits of increasing trial number are visible when there were 5 trials per condition (middle panels). At the largest effect size, the design would nearly always detect the main effect of condition. However, this design is still quite insensitive to the real individual differences I included in the simulation, and exact replication studies would still only return significant results around 50% of the time for the 100ms effect. This contrasts with the 100 trials per condition design (rightmost panels), which, in addition to detecting even the smallest effect very reliably, was also sensitive to many of the individual differences between animals within each study simulation. Studies using such designs in animal cognition can produce, and in many areas already do produce, very replicable results.

However, this benefit of running many trials across participants may not be accessible for some areas of animal cognition. This is the case when theoretical constraints restrict experimental designs to

using a single test trial, in order to avoid experimental confounds. For example, in some memory tests the animals are tested in extinction, to prevent cues from the to-be-remembered items directly influencing their searching or responding behavior (Adams & Dickinson, 1981; Clayton & Dickinson, 1998). Such studies cannot use many test trials in within-subjects designs because this would invalidate the test itself; for example, the animals could learn that no food is provided in the test trials and therefore that there is no value in searching for the food. Hence, for some hypotheses in animal cognition, the optimal methodological design may not be one that produces highly replicable results, if replicability is measured by statistical significance only. Nevertheless, all else being equal, studies employing many trials will produce more replicable results than studies using few trials. However, this does not mean that most many-trial studies and no few-trial studies will replicate successfully. Further, even if this were true, it does not mean that many trial studies are always "better". As ever, there are trade-offs between likely replicability, validity and resource investment. Numerous experiments that use many trials, such as reaction time studies in humans, can have participants perform hundreds of trials in minutes, whereas in many animal studies this is not possible. The reliability-resource trade-off is just that: a trade-off, and researchers should seek to evaluate and improve reliability with this in mind.

Conclusion 3: Some areas of animal cognition face replicability/validity and replicability/resource trade-offs

2.1.3.3. Replications across animal cognition: Small-N, Many Replicates?

One research area that consistently uses many trials to produce replicable results is so-called "small-N research". Small-N research capitalises on the benefits of using many trials on each individual subject, yet rather than seek to estimate population parameters they treat the individual as the replication unit (Little & Smith, 2018; Smith & Little, 2018). Each individual is its own experiment, and experiments are thus replicated by using more than one individual. Smith and Little explain this approach:

"We argue that some of the most robust, valuable, and enduring findings in psychology were obtained, not using statistical inference on large samples, but using small-N designs in which a large number of observations are made on a relatively small number of experimental participants. We argue that, if psychology is to be a mature quantitative science, its primary theoretical aim should be to investigate systematic, functional relationships as they are manifested at the individual participant level. The estimation of population parameters, while not unimportant, is arguably of secondary concern and should probably be investigated using more refined techniques for characterizing individual differences than the blunt instrument of simple averaging that conventional statistical methods provide."

Smith & Little, (2018, p. 2084)

The small-N approach is found across psychological research, for example in Ebbinghaus' forgetting curve and many psychophysical experiments. And interesting hypotheses are often tested at the individual level in animal cognition - think of studies of number comprehension in Alex the parrot (Pepperberg & Gordon, 2005), of fast mapping with Chaser the dog (Kaminski et al., 2004) and of working memory in Ai and Ayumu (Inoue & Matsuzawa, 2007). Individual-level data are also widely used in comparative psychology. Learning curves, for example, are usually most informative when plotted at the individual level, and this is part of the reason why Skinner rejected null-hypothesis significance testing (Gigerenzer et al., 2004; Skinner, 1956). Small-N research and individual-level data are particularly informative when we can be confident that the results will generalize to at least some more individuals. For example, understanding the learning mechanisms of Lab Rat 377 logically will inform us about rat learning more generally, and hence a small-N approach might be appropriate here. However, while small-N research has a formal approach to inference at the individual level (P. L. Smith & Little, 2018), there is no formal method of generalizability. In animal cognition, when strong evidence is found at the individual level, the question "will it replicate?" asks "how will it generalize to similar individuals who have had similar experiences?". In contrast, group-level animal cognition research, the question "will it replicate?" may be more synonymous to "is it a reliable statistical effect in the given population?", but see Chapter 3 for a deeper discussion.

Conclusion 4: Different research approaches may have different meanings by replicability (see e.g., Lazic et al., 2018). While these approaches answer different questions, animal cognition can benefit from both approaches, and can often employ both simultaneously.

2.1.3.4. The data are not everything

Thus far, I have considered only statistical and design features of replicability. However, there is much more information about replicability than can be gleaned from reported statistics alone, especially if there is a publication bias. When Daryl Bem provided evidence for a physically impossible skill, precognition, across nine studies, statistical markers were part of the subsequent refutation (Francis, 2012; Schimmack, 2012; Wagenmakers et al., 2011). However, it was the sheer implausibility of the effect that provided the clearest indicator that it would not replicate (Chambers, 2017). This ability of researchers to detect findings that are likely unreliable extends from the physically impossible to more plausible research results too. Research using forecasting and prediction markets has shown that groups

of experts, on average, can be surprisingly accurate at identifying research that will not replicate (Dreber et al., 2015; Forsell et al., 2018) – and there is no reason why this should be any different in animal cognition research for the statistical reliability of effects. In animal cognition research, if your belief in something is very low, then a just-significant result should not affect this belief greatly. This is not to say that researchers should be insensitive to evidence, but they should critically assess whether there is sufficient evidence that a given statistical effect is accurate, and whether this statistical effect is strong evidence of the authors' claims. Extra-ordinary claims require extra-ordinary evidence, and this will usually not be provided by single studies with just-significant results, or even a set of such studies if this set of studies cannot provide an effective risk of bias assessment.

Conclusion 5: The data aren't the full story – aggregated expert beliefs about replicability might be accurate, whatever they are based on

2.1.3.5. Difficulty performing and interpreting replication studies in animal cognition

Given that many research findings might be unreliable in animal cognition, a next step could be to call for a suite of systematic, direct replications studies to identify findings that are robust, which would enable researchers to perform meaningful meta-analyses on rich datasets with low bias. While there are strong reasons to support such claims (e.g. see Beran, 2018; Lambert et al., 2021; Many Primates, Altschul, Beran, Bohn, Call, et al., 2019; Stevens, 2017), there are many valid barriers to performing and interpreting replication studies in animal cognition that should be considered. These barriers mean that it may not be possible to perform truly direct replication studies in most animal cognition research, because:

- Restricted resources mean that it is not possible to directly replicate some findings
- Statistical estimates from both original and replication studies will be too noisy to be able to detect differences between them with confidence
- There will be real and often large differences between animal behaviour in original and replication studies

Restricted resources mean that it is not possible to directly replicate some findings

As animal cognition is a small field represented by many different research questions in many different species (Beran et al., 2014; Shettleworth, 2009), when a laboratory stops working on a certain species the likelihood of direct replication studies of these results in the near future approaches zero. For example, our Animal cognition laboratory in Cambridge performed a series of studies on cache-protection strategies in California scrub-jays (*Aphelocoma californica*) which it no longer houses. Consequently, the

possibility of these studies being directly replicated with California scrub-jays in this lab in the coming years is very low. Multi-lab collaboration, or independent replication can help, for example there is one laboratory that currently publishes on caching in California scrub jays (Clary et al., 2019). However, it is unclear to what extent different labs are incentivised to replicate others' findings in animal cognition, especially given their own funding and ethical constraints.

Conclusion 6: Often, direct replications will not be immediately possible due to a lack of resources and researchers should explore collaborative methods to estimate the reliability of their research.

Statistical estimates from both original and replication studies will be too noisy to be able to detect differences between them with confidence

In order to detect a difference between the results of an original study and its direct replication, only looking at the significance of the results is not appropriate. Considering statistically significant replications to be the only "successful" replication studies, and all others to be "unsuccessful", is misleading (Gelman, 2018) as it dichotomizes replication attempts as either successful ($p < 0.05$), or unsuccessful ($p > 0.05$). While a replication study with $p = .049$ would be considered a success, a replication study with $p = .051$ would be considered a failure. If the sample size of a replication study is not much larger than the original study, it is not surprising that many replication studies will not yield significant results, even if there is a real effect being studied. In fact, more liberal, and perhaps more appropriate, interpretations of large-scale replication studies results provide higher estimates of replicability than those first publicized. For example, it is often reported that the Open Science Collaboration (2015) produced only 35 positive replication results from 97 positive original findings. However, Etz and Vandekerckhove (2016, p. 1) used a Bayesian analysis to suggest that "75% of [replication] studies gave qualitatively similar results [to the original studies]", but also noted that "the majority of the studies (64%) did not provide strong evidence for either the null or the alternative hypothesis in either the original or the replication." Future replication projects in human psychology have attempted to address this concern by swamping the sample sizes of the original studies (e.g., Camerer et al., 2016). However, as Morey and Lakens note (2016, p. 1), "sample sizes are so small in psychology that often one cannot detect even large differences between studies. High-powered replications cannot answer this problem, because the power to find differences in results from a previous study is limited by the sample size in the original study". When this concern is translated to animal cognition, and its constraints on sample sizes, it is clear that producing replication studies that can assess the veracity of the original claim will be very challenging.

Conclusion 7: Animal cognition might not have the resources to produce highly-powered and informative replication studies of many claims

Real and large differences between animal behavior in original and replication studies

Assuming that some form of direct replication studies are performed, interpreting their results becomes even more difficult when the many reasons why a animal cognition replication study might fail to produce significant results are considered, even if the original effect was "true". One example is that animal behaviour often has large seasonal and developmental variation. For example, a food-caching experiment would fail to replicate outside of caching seasons, and a memory experiment performed on young animals might not replicate when the same animals are tested in their old age. Such failures to replicate could be due either to the original results being a false positive, or a well-motivated alternative hypothesis, like the memory performance of animals decreasing with age. Furthermore, the experiences of animals, particularly those that are highly trained on certain apparatuses, can prevent results from being easily replicable. This can be either in principle, e.g. some zoos might not be able to house the same equipment of laboratories, or in practice, e.g. the same equipment might have different effects in different laboratories.

Conclusion 8: Temporal and developmental variation in animal behavior will influence the likelihood of replication success

Needless to say, when direct replication studies using a new sample from the same population can be performed, these will be an effective method of assessing the reliability of an effect. However, in animal cognition, it is not feasible to assume new samples can be repeatedly taken from the same population – in fact it is often unclear what "populations" we are studying in general. Rather, different groups of animals of the same species from different research sites may be best viewed as different populations with respect to many cognitive effects. In biomedical research, researchers report differences in physiology and behaviour between laboratories, even when they test the exact same strain of animals with the exact same protocol Crabbe et al., 1999). These real differences in nominally similar animals when exposed to the same treatment means that results are difficult to reproduce between-laboratories (Voelkl & Würbel, 2016)[3], but also within laboratories when the same animals are tested repeatedly (Karp

[3] These differences in behaviour might reflect an adaptive response to many subtle environmental differences between sites, and as such "independent replicate studies that fail to reproduce the original findings might not necessarily indicate that the original study was poorly done or reported, but rather that the replicate study was probing a different region of the norm of reaction." (Voelkl & Würbel, 2019, p. 3) The reaction norm is a concept

et al., 2014). That these differences are present in highly standardized conditions, such as biomedical mouse research, raises the likelihood that real and potentially large effects of site and time will reduce the replicability of animal cognition research.

A recent example in animal cognition comes from the ManyPrimates collaboration (Many Primates, Altschul, Beran, Bohn, Call, et al., 2019; Many Primates, Altschul, Beran, Bohn, Caspar, et al., 2019). Here, they collected data on 176 individual primates from 12 species on a delayed response task, in which primates had to wait either 0, 15 or 30 seconds before choosing the location where food was hidden (Many Primates, Altschul, Beran, Bohn, Call, et al., 2019). While there are notable similarities within and between species' behaviour across sites, some large differences can still be observed. For example, the 12 chimpanzees from the Wolfgang Köhler Primate Research Center greatly outperformed the 12 chimpanzees from Edinburgh Zoo. If such a between-site difference is present in a relatively simple and robust cognitive task, in which inter-individual differences might be expected to be low (Hedge et al., 2018), this suggests that even larger between site differences could manifest for more noisy behaviours in animal cognition replication studies. While in the ManyPrimates data the variation within chimpanzee performance was low compared to the variation between species, and some species, e.g. capuchin monkeys, were very similar across sites, it was only possible to know this because they did sample different groups of animals from the same and different species. In contrast, multi-site studies in animal cognition that do not sample from multiple groups of the same species risk confounding or obscuring between-species differences in behaviour because they are unable to dissociate the contributions of species and site differences to the data. Even when making within-species comparisons, direct replications between sites in animal cognition could be seen as similar to cross-cultural studies in humans, and as such lie closer to conceptual rather than direct replications. Chapter 3 focuses on this topic.

Conclusion 9: Site specific differences in behaviors make direct replication studies sampling from new populations difficult to interpret, and could confound many between species comparisons

Overall, it will not be possible for researchers to identify results as "false positives" through a limited number of direct replication studies. Often, a researcher wanting to replicate a study will not have access to the desired species, and even if they did, they would be unlikely to be able to produce a result that they could confidently conclude is statistically different from the original study. Furthermore, our default assumption should be that there will be real quantitative differences between the results of

used by Voelkl and Würbel to represent gene x environment interactions acknowledging both plasticity and canalization.

replication and original studies. This holds both for effects which we might be confident are true positives, for example the observed difference between Edinburgh Zoo chimpanzees and Wolfgang Köhler Primate Research Center chimpanzees in ManyPrimates, and for effects that we might be uncertain are true, such as those supported by a few just significant *p*-values. Many of the claims that are supported by a few just significant results may be practically unfalsifiable – precisely because we would not expect them to replicate consistently in similar studies even if there was a true effect. However, this does not mean that all findings reflect "true effects" in animal cognition, instead it means that single small replication studies will not absolutely falsify a claim (but similarly neither have many original studies proven their claims). Instead of focusing on the "presence" or "absence" of an effect in a replication study, a more fruitful approach would be to focus on whether there are meaningful differences between original and replication (Gelman, 2018; Morey & Lakens, 2016). However, often the estimates from both the original study and the replication study will be so imprecise that we often will not have the resources to detect these differences. To account for this, animal cognition researchers can focus on expressing the uncertainty about their findings, rather than the absolute veracity of a claim or its statistical significance. This should act prospectively when interpreting the results of replication studies and new studies, but also retrospectively for small studies that have made bold claims on the basis of weak statistical evidence.

Conclusion 10: Single replication studies are unlikley to absolutely falsify claims. A focus on effect sizes, meaningful differences between studies and communicating uncertainty should be a long-term aim for animal cognition research.

2.1.4. The beginning of a replication crisis in animal cognition?

Over the last few years, replication has received increasing attention in animal cognition research (Beran, 2020; Brecht et al., 2021; Freeberg, 2020; E. Tecwyn, 2021), and the predictions made in this chapter were somewhat borne out across published replication studies. To illustrate this, I highlight four recent replication "failures" from avian cognition research, two from my own lab using caching studies to examine Eurasian jays' cache protection strategies at the group level (Amodio, Farrar, et al., 2021; Crosby, 2019), a study of New Caledonian crow physical cognition (O'Neill et al., 2021), and an individual-level research project on mirror recognition (Soler et al., 2020). Table 4 highlights these studies, along with the authors' interpretations.

Table 4: Four recent claimed replication failures in avian cognition research

Replication study	Original study/studies	Interpretation
Amodio, Farrar, et al. 2021	"Eurasian jays (Garrulus glandarius) conceal caches from onlookers" Legg and Clayton, 2014 "Current desires of conspecific observers affect cache-protection strategies in California scrub-jays and Eurasian jays" Ostojić et al., 2017	"Experiments 3, 4 and 5 found no significant effects in the direction of the previously reported effects, questioning their robustness…. …We propose two explanations for why our studies were unable to detect effects consistent with the previous literature, namely low power and the re-use of a unique bird sample" p.20 of accepted manuscript
Crosby 2019	"Current desires of conspecific observers affect cache-protection strategies in California scrub-jays and Eurasian jays" Ostojić et al., 2017	"However, I was unable to replicate the results of the original study, which may lead to questions about the reliability of this effect. As described above, there are two possible explanations for the inability to find an effect here: (i) the replication was a false negative, or (ii) the original result was a false positive." p. 107
Soler et al. 2020	"Mirror-Induced Behavior in the Magpie (Pica pica): Evidence of Self-Recognition" Prior et al., 2008	"Thus, our replication failed to confirm the previous results. Close replications, while not disproving an earlier study, identify results that should be considered with caution." p. 363
O'Neill et al. 2020	"New Caledonian crows reason about hidden causal agents" Taylor et al., 2012	"The low sample size of our replication group meant we could not be sure if we did not replicate the effect due to low power or due to actual differences." p. 166

All four replication studies from Table 4 followed the same pattern: an attempt to match the methods of a previous study as closely as possible with a similar sized sample to an original study. Each original study produced a statistically significant result a large effect sizes. These replication attempts all returned non-significant results, and the authors were unable to conclude anything other than that their findings question the robustness of the previous findings. This is an appropriate interpretation – the risk of false negative results along with the wide confidence intervals around both the original findings and replication results mean that the replication studies cannot disprove the original findings. This is where

the replication crisis in animal cognition will likely differ from, e.g., human social psychology. In human social psychology, it is often possible to generate large sample sizes and have relatively few constraints on how tasks can be implemented (e.g., Camerer et al., 2018), i.e., it should be possible for many low-cost studies to perform replication studies that do not just match original studies, but are unequivocally better studies than the originals. In animal cognition, this will often be infeasible (although this may be similar to other resource intensive fields with difficult-to-reach populations, such as studies of rare diseases Lange (2019)). Hence, the replication "crisis" in animal cognition may not be that many effects are demonstrated to be unequivocally false, but the realisation that we have literatures filled with research performed under conditions that promote false positive findings (Chapter 4).Well-performed replication and extension studies are an important tool to test the generalisability of these published research findings, but this begs the question of what a replication and what a "well-performed" replication is in animal cognition. Chapter 3 now attempts to develop a deeper understanding of what a replication is, and how this relates to theory testing in animal cognition.

3. Chapter 3: Replications, sampling, and theory testing[4]

Chapter 2 outlined reasons why animal cognition research might expect a low rate of successful replication across many of its studies. In addition to biasing factors, such as publication bias or *p*-hacking, an important consideration was that animal cognition research often involves small and idiosyncratic samples, that can vary over time. When a new sample is taken, this sample is sufficiently different from the original that we should expect there to be some real differences between the original and replication results. However, when researchers make general claims about a group or species' behaviour, they implicitly assume that their samples are representative enough of the group or species of interest. Yet, this assumption is rarely tested, and the literature is populated by claims that are produced by single laboratories, testing the same animals, at single time points and in closely related experimental designs. This could lead to overgeneralised findings that are difficult to replicate (Henrich et al., 2010; Würbel, 2000; Yarkoni, 2019), but equally, it could be an effective strategy to maximize scientific progress in resource-limited fields (Craig & Abramson, 2018; Davies & Gray, 2015; Mook, 1983; Schank & Koehnle, 2009; Smith & Little, 2018). To explore this issue, this chapter shows how concerns about replicability, representativeness, comparison and theory testing and pseudoreplication are all related through the lens of sampling. To design the best experiments, researchers should consider all five in relation to their sampling plans. The first half of the chapter focuses on sampling and replication, and answers the following questions:

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

The second half of the chapter then focuses on representativeness and asks how concerned researchers should be with the problem of non-representative sampling in animal research. I explore this issue through a re-analysis of existing data on animal "self-control" and a simulation study. The simulation study shows that for some between-group or between-species comparisons, poorly representative samples could lead to false positive rates closer to 50% than 5%, the rate conventionally cited when

[4] This chapter contains material published in Farrar, B. G., Voudouris, K., & Clayton, N. S. (2021). Replications, Comparisons, Sampling and the Problem of Representativeness in Animal Behavior and Cognition Research. Animal Behavior and Cognition, 8 (2), 273-295

authors use $p < .05$ to define statistical significance. Finally, the chapter ends with a discussion of how researchers might assess, mitigate and account for the problem of representativeness in animal cognition.

3.1. Claims, Samples and Replications

3.1.1. What are replications in animal research?

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

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

Recently, resampling definitions of replication have been developed (Asendorpf et al., 2013; Machery, 2020). These may be the most effective definition of replication in animal cognition research. When researchers test a claim, they sample from populations of experimental units (most often animals), settings, treatments and measurements (Gómez et al., 2010). For example, when testing the claim that

chimpanzees will explore a mark on their forehead when exposed to a mirror, researchers sample from the population of chimpanzees available for research, from various settings (laboratories, zoos, wild), with a variety of possible treatments (different size mirrors, different types of marks, etc.), and many different possible measurements (e.g., an ethogram of self-directed actions). The resampling definition of replication states that a replication study is: a study that resamples from the same populations of experimental units, treatments, measurements and settings that an original study could have sampled from, relative to the claim being tested (Machery, 2020; Nosek & Errington, 2020). This is outlined in Table 5, adapted from Machery (2020).

Table 5: A Resampling Account of Replication (Adapted from Machery, 2020)

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

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

To see how the resampling definition can be applied in animal cognition research, I now discuss a partial or “conceptual” replication of a study investigating ageing in monkeys (Almeling et al., 2016; Bliss-Moreau & Baxter, 2019). This is a useful example as, like most experiments in animal cognition, Bliss-

Moreau and Baxter's study is not a close replication of the previous study; it was neither conducted in identical laboratory settings nor even in the same model species.

3.1.2. Case Study: Do Non-Human Primates Lose Interest in the Non-Social World With Age?

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

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

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

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

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

One barrier to identifying and testing claims is that many theories and claims in animal cognition are verbal and vague (Bourjade et al., 2020; Farrar & Ostojic, 2019). This makes it difficult to derive risky predictions of the theories, because their vagueness affords them the flexibility to accommodate nearly any result (Roberts & Pashler, 2000). This could be remedied by formally modelling theories and hypotheses (Farrell & Lewandowsky, 2010; Guest & Martin, 2020). Such models may be key to making progress in understanding animal minds (Allen, 2014), and they can be informed by known mechanisms driving animal behaviour, such as associative learning (Heyes & Dickinson, 1990; Lind, 2018; Lind et al., 2019). However, these models need not be preferred, or even contradict, non-associative models (Bausman & Halina, 2018; Mercado, 2016; Smith et al., 2016). Just like any other scientific tool, formal models need critique from a variety of perspectives, and this is developed in Chapter 4.

3.2. Species-Fair Comparisons

The resampling account not only offers a theoretical framework for replications, generalisations and theory testing in animal cognition research, but it also offers a framework for analysing between-species comparisons. Between-species comparisons are just tests of the generalisability of an effect across species, and like any other test of generalisation they can be reframed in terms of replication, too. Comparing an effect between a group of chimpanzees and a group of bonobos is the same as testing if the effect generalises from chimpanzees to bonobos, or replicating a study performed in great apes, albeit with two non-random and systematically different samples. Both of these experiments would be entailed by the coarser question of whether great apes (chimpanzees, bonobos and orangutans) show the effect

in question. Whether the study in question is best described as a comparison, replication or a test of a claim is somewhat moot — it is all three at the same time, relative to claims of different coarseness.

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

3.3. The Problem of Representativeness in Animal Research

A sampling perspective shines light on why many results in animal research may struggle to replicate. Animal experiments often sample a small number of animals at a single site, using a single apparatus and measurement technique. However, from these small samples come general claims about animal behaviour, creating a mismatch between the statistical model and the theoretical claim (Meehl, 1990; Yarkoni, 2019). The statistical model will usually allow generalisation to the population that the experimental units were randomly sampled from, for example the population of animals at a given site, (although even then they may not be randomly sampled, see Schubiger et al., (Schubiger et al., 2019), but any inferences to the wider population of interest will be overconfident, unless the population of interest can be justified as the individual animal (Lazic et al., 2018; Smith & Little, 2018). This is an unavoidable consequence of working with difficult to reach populations (Lange, 2019), but it should be accounted for when building theories. This is important as animal behaviour does seem to vary across samples, for example due to experimenter effects (Beran, 2012; Boesch, 2021; Bohlen et al., 2014; Cibulski et al., 2014; Pfungst, 1911; Sorge et al., 2014), genetic variation (Fawcett et al., 2014; Johnson et al., 2015; MacLean

et al., 2019), housing conditions (Farmer et al., 2019; Hemmer et al., 2019; Würbel, 2001), diets (Davidson et al., 2018; Höttges et al., 2019), and of course learning/developmental histories (Skinner, 1976).

3.3.1. Situating the Problem of Representativeness

If researchers are wary of just how much animal samples can vary due to factors often labelled as "noise", then they might be concerned about the representativeness of the samples they test (or of themselves while they are testing, or the site they are testing at). This problem of representativeness has been discussed from several different angles across scientific literatures, albeit often with different terminologies and little connection between them. However, they share the similar underlying concern that researchers' claims are poorly matched by their sampling strategies and statistical models, and I outline these briefly now:

3.3.2. Replicability

First, a lack of representative sampling causes low replicability. Because of small and non-representative samples of experimental units, settings, treatments and measurements, sampling variation will mean that laboratories will struggle to replicate or reproduce the results of previous experiments. This argument has featured heavily in rodent phenotyping studies (Crabbe et al., 1999; Kafkafi et al., 2005, 2017, 2018; Lewejohann et al., 2006; Richter et al., 2009, 2011; Wahlsten et al., 2003; Würbel, 2000).

3.3.3. Generalisability and External Validity

Second, a lack of representative sampling causes problems of generalisability or external validity: researchers claims will not often generalise to novel but related settings (Mook, 1983; Yarkoni, 2019).

3.3.4. Pseudoreplication

Third, the lack of representative sampling in animal research is usually due to non-random sampling from the population of interest. This leads to pseudoreplication (Hurlbert, 1984; Lazic, 2010; Waller et al., 2013), if this non-random sampling is not accounted for in the statistical models, and the consequence is that uncertainty intervals will be overly narrow, and the results will struggle to replicate in new samples – or generalise to them.

3.3.5. Theory Testing

Fourth, the lack of representative sampling produces weak tests of a theory or claim (Baribault et al., 2018; Nosek & Errington, 2020): a test that probes only a small sample space of a theory's predictions

provides less opportunity for weaknesses in the theory or claim to be found, compared to a test which covers most of the relevant sample space.

3.4. The Difficulty of Identifying the Sources of Differences Between Groups and Species

That animal behaviour differs across space and time makes it difficult to understand whether species or group differences in behaviour are really a consequence of theoretically interesting species or group differences, or whether they are due to the host of other factors that vary between sites. In reality, the observed differences between two groups will be the sum of the group differences in behaviour that are of theoretical interest and all other factors that influence animal behaviour and vary between sites. When making quantitative between-species and between-group comparisons, they are nearly always confounded by site-specific differences in factors that are not the focus of interest. Lazic (2016) commented on such a scenario in an introductory textbook for laboratory biology:

"To make valid inferences, one would need to assume that the effects of [site] are zero. Moreover, as this assumption cannot be checked, the researcher can only hope that [site] effects are absent. Such a design should be avoided." (Lazic, 2016, p. 68)

One may object to this and acknowledge that while there are many variables that differ between sites but go unmeasured, the net sum of these effects should be close to zero across sites, i.e., they will cancel each other out. However, this would only be the case if there were many variables with small effect that were randomly assigned to each site, and this is not what happens. On the contrary, laboratories or sites differ markedly from each other on a range of variables with large effects (e.g., housing conditions, learning experiences). It is often recognised that animal laboratories are poorly positioned to generate representative data of the species in the wild (Boesch, 2021; Calisi & Bentley, 2009), but what if they are also poorly positioned to generate representative data of the species in laboratories? Taken to the extreme, there may be a laboratory testing a sample that is more representative of a species other than its own, for example a sample of lemurs that have parrot-like self-control, or a sample of hand-reared wolves that behave more like dolphins when presented with a novel object. To highlight the difficulties of making between group or between species inferences across sites, I now present a case study of between species comparisons made using the cylinder task, and then present a simulation study of how sampling affects comparisons in animal research.

3.5. Case Study: Between Species Comparisons and the Cylinder Task

For this case study I used data from MacLean et al. (2014) to probe the stability of a measurement of behavioural inhibition when taking new samples of experimental units at new sites. MacLean et al.'s (2014) large-scale study tested the performance of 36 species across 43 sites on two tasks aimed at measuring self-control (but rather measured one form of behavioural inhibition: Beran, 2015); the A not B task and the cylinder task. The cylinder task was given to 32 species across 38 sites. In this task, animals are familiarised with retrieving a piece of food from the centre of an opaque cylinder. After retrieving the food from the opaque cylinder in 4 out of 5 consecutive training trials, the animals proceed to testing. In testing, the animal is presented with a transparent cylinder with food in the centre. In order to successfully retrieve the food, the animal needs to inhibit an initial drive to directly reach for the food and subsequently collide with the transparent cylinder, and instead detour to the cylinders ends to access the food. Each animal was given 10 trials, and an overall score between 0% (no animals succeeded on any trial) and 100% (all animals succeeded on every trial) was computed for each species. Five species (orangutans, gorillas, capuchin monkeys, squirrel monkeys and domestic dogs) were tested across two sites. Figure *2* displays the between-site variation for these samples, and also includes data from a sample of an additional species, the Western scrub-jay, that was tested both in the original experiment and a couple of years later at a new site (Stow et al., 2018). Figure 2 displays these data, with the samples performing close to ceiling towards the top of the figure and those that did not perform close to ceiling at the bottom.

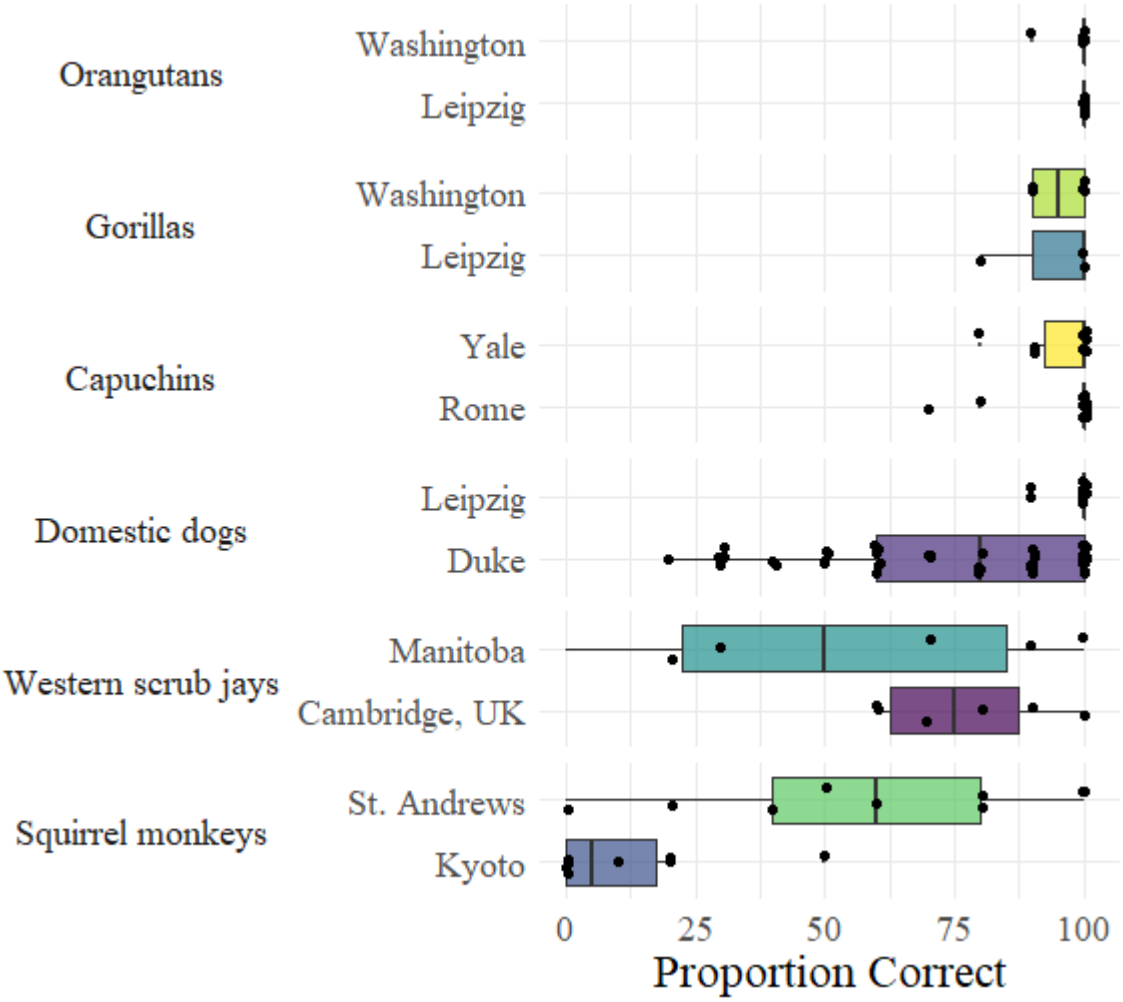


Figure 2: Species Differences Between Sites in the Cylinder Task. All data from MacLean et al. (2014), except the Manitoba scrub-jay data, which are from Stow et al. (2018).

For the samples of scrub-jays, squirrel monkeys and domestic dogs, this variability is striking. For squirrel monkeys, the median score in Kyoto was 5%, compared with 60% in St Andrews. No individual in Kyoto performed above the median in St Andrews, and this demonstrates how some between-site differences that cannot be attributed to species identity can have large influences on behaviour. To highlight the issues this can pose for inference, consider what would happen if the animals from Kyoto were not squirrel monkeys, but Tonkean macaques. Then, it is likely that the difference in performance compared to the St Andrews' squirrel monkeys would likely be interpreted as a species difference – "Tonkean macaques are worse at behavioural inhibition than squirrel monkeys", could be the title of a paper reporting these results. In fact, the substantial difference in behaviour between species tested at different sites need not imply meaningful species differences at all. If we took new samples for all species that MacLean et al. tested, it is possible a completely different ranking of animals would be produced.

MacLean et al.'s (2014) overall model gains credibility, however, because of the use of phylogenetic models (and also including data from the A-not-B task, another test of behavioural inhibition). Incorporating phylogeny and estimating phylogenetic signal when making comparisons, providing there is enough data, little bias, and sufficient model checks, can lead to large increases in statistical power (Freckleton 2009; see MacLean et al., 2012 also for an overview of other benefits of comparative phylogenetic models). However, any individual site comparison of non-ceiling cylinder task performance between species, either within the MacLean et al. study, or from other published research, is likely too uncertain to produce meaningful estimates at the species level, and this can lead researchers astray when making inferences from individual results. Table 6 presents some statements from studies that followed MacLean et al.'s procedures using a single sample of a species at a single site, along with the species' cylinder task "score":

Table 6: Results and Claims from Samples of Four Species Tested on the Cylinder Task

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

This set of numerical comparisons are factually correct, but what do they mean? The worst performing chickens actually scored higher than the Kyoto squirrel monkeys, and if we sampled another population of great tits it is possible that their performance would regress close to the mean value of all species. Ordaining a species with a single score following a single test on a small sample of animals from

a single site with a single apparatus, and then comparing this number between species has no means of error control and hides the uncertainty in their estimates. Several of the inferences are reasonable, for example we may genuinely believe that chickens will perform poorly on behavioural inhibition tasks, but this is primarily constrained by our prior beliefs about chicken cognition.[5] For potentially more surprising results, such as the high score of great tits, our beliefs are not so constraining, yet neither are the data.

Moreover, and counter-intuitively, the best estimate of great tit performance on the cylinder task is not the 80% reported by Isaksson et al., even though this is the only known data collected with great tits on this task. Rather, the best estimate would utilise the information we have about similar animals (other birds of a similar size/socio-ecology/phylogeny), that would shrink our estimate of great tit performance closer to the mean value for, as an example, all Passeriformes tested to date. Interestingly, during the revision process of this article, three further datasets of great tit performance on the cylinder task became available. In contrast to the 80% reported by Isaksson, and in line with our prediction of regression to the mean, Troisi et al. (2020) recorded a score of 38%, and a sample of 35 tested by Coomes et al. (2020) scored 41%. Moreover, in a pilot to one of these studies using a larger tube, a sample of great tits scored 0%, suggesting that the size of the tube can heavily modulate individual's performance (G.L. Davidson, personal communication). While there were differences in the experience of Iskasson et al.'s birds, who had some previous experience with transparent cylinders, it is not clear that such a difference can account for the higher scores of these birds without also considering sampling bias.

How, then, can we make better inferences from single site samples of data? We could either attempt to get a better estimate at this single site, for example by testing great tits on a wide range of tube apparatuses. Alternatively, we can also use the data from other species to inform our great tit estimate. Because the behaviour of different animals will often be correlated, for example as a function of phylogenetic distance, we should allow data from similar species to guide each other's estimates. Ideally, a phylogenetic model would be constructed which incorporates information on the phylogenetic distance between species and a model of the trait's evolution (McElreath, 2016). Other relevant predictor variables, such as body size, tube size or body size/tube size ratio, could also be added into these models, or they could be investigated in separate meta-regression models. However, for many animal cognition questions such models will be difficult to generate, but the general principle holds: when a surprisingly

[5] During my viva I had labelled this belief as "arbitrary". Marta Halina rose the question about what makes a belief this like arbitrary, and how can prior beliefs like these be evaluated, especially if they differ across research groups or experts. I had no good answer for this, but I think survey studies or Delphi panels to record animal cognition researchers' actual beliefs on certain topics, and how they update, would be an interesting first step.

high or surprisingly low estimate of a species behaviour is produced, and most data from similar species are less extreme, it is likely that the new estimate is over- or underestimated. Returning to the cylinder task, it is clear that non-ceiling results are not very informative about animal cognition if we do not know whether the results from any given sample are stable across space or time — before considering issues of construct validity (Beran, 2015; Kabadayi et al., 2017, 2018).

3.6. Programme-wide confounds

The confounding effects of experimental units, measurements, treatments and settings that I have considered so far may not be a priori strong candidates as confounding variables. For example, there is no single obvious reason why Manitoba scrub-jays might perform differently to Cambridge scrub-jays (but see Stow et al., (2018) for some suggestions). It is also possible that we may never know what they are – we won't be able to gather enough data (Blastland, 2019; Crabbe et al., 1999; Voelkl & Würbel, 2019). But in some cases, it is possible to pinpoint likely confounders that are near perfectly correlated with species, and this can be used to reduce confidence in whole bodies of data. For example, Clark and colleagues (Clark et al., 2019; Clark & Leavens, 2019) highlighted how procedural differences in an object choice task confound species difference inferences between dogs and non-human primates, i.e., the comparisons lack validity because the measurements and settings may be from different relative populations between dogs and primates. Clark et al. found that over 99% of non-human primates tested on the object choice task were tested with a barrier between the experimenter and participant, whereas this was the case for less than 1% of the dogs. When the dogs were tested behind the barrier, their performance decreased (Kirchhofer et al., 2012), suggesting that any difference between dogs and nonhuman primates is at best overestimated. Similarly, Boesch (Boesch, 2007) and Leavens et al. (2019) provide clear arguments appealing to such between species confounds in ape-human comparisons: "all direct ape–human comparisons that have reported human superiority in cognitive function have universally failed to match the groups on testing environment, test preparation, sampling protocols, and test procedures, including those that tested subjects' comprehension and production of communicative gestures" (Leavens et al., 2019, p. 491) Clark et al. and Leavens et al. provide archetypal failures of comparison: different experimental units being tested using relatively different apparatuses in relatively different settings, across entire research programmes. However, this argument can be made for most between-species comparisons in disciplines that compare groups of animals' performances at different sites. For most comparisons, the question is not, "are they confounded?", but "what are the consequences of the (un)known confounding variables?".

3.7. Simulation Study

To illustrate how between-site variation (a proxy for the sum of setting, treatment and measurement variation) can lead to elevated false positive rates and results that struggle to replicate, I now present a short simulation study of a replication and a comparison in animal cognition. The simulation and visualizations were performed in R 4.0.2 (R Core Team, 2020), using the packages *tidyverse* (Wickham et al., 2019), *extrafont* 0.17 (Chang, 2017) and *scales* 1.1.1 (Wickham and Seidel, 2020). The code is available at: https://github.com/BGFarrar/Replications-Comparisons-and-Sampling. This section can be skipped if the reader is already comfortable with the topic. I simulated a hypothetical within-species replication, between two groups of chimpanzees, and a hypothetical between-species replication/comparisons between a group of chimpanzees and a group of bonobos. I simulated 100 hypothetical sites of chimpanzees, and 100 hypothetical sites of bonobos, with 100 animals at each site. The behaviour of animals within a site was correlated, such that animals sampled from the same sites, on average, had more similar behaviours than animals sampled from different sites. At each site, I "measured" each animal's behaviour to produce a neophobia and self-control score for each. For both the replication simulation and the comparison simulation, four parameters were used to simulate each animal's behaviour: a population grand mean, β_0, a by-location random intercept $L_{0,}$ a by-subject random intercept $S_{0,}$ and a by-individual residual error term e_{ls}. Subject was nested within location, such that all subjects at the same location had the same location effect. Data were simulated using the following formula:

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

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

Neophobia

β_0 = 800

L_{0l} ~ *N*(0, 100)

S_{0s} ~ *N*(0, 100)

e_{ls} ~ *N*(0, 50)

Self-control

β_0 = 80

L_{0l} ~ *N*(0, 10)

$$S_{0s} \sim N(0, 10)$$

$$e_{ls} \sim N(0, 5)$$

The panels of Figure 3 display the behaviour of all 10,000 chimpanzees (100 animals x 100 sites) in grey. Next, I randomly selected one site to be our first sample. The upper panel of Figure 3 highlights all 100 chimpanzees from this site. However, in reality we would not usually have access to or test 100 animals at a site, instead a primate cognition sample size is usually around 7 (Many Primates, Altschul, Beran, Bohn, Caspar, et al., 2019). Therefore, I randomly selected 10 animals, which are highlighted in the lower panel of Figure 3.

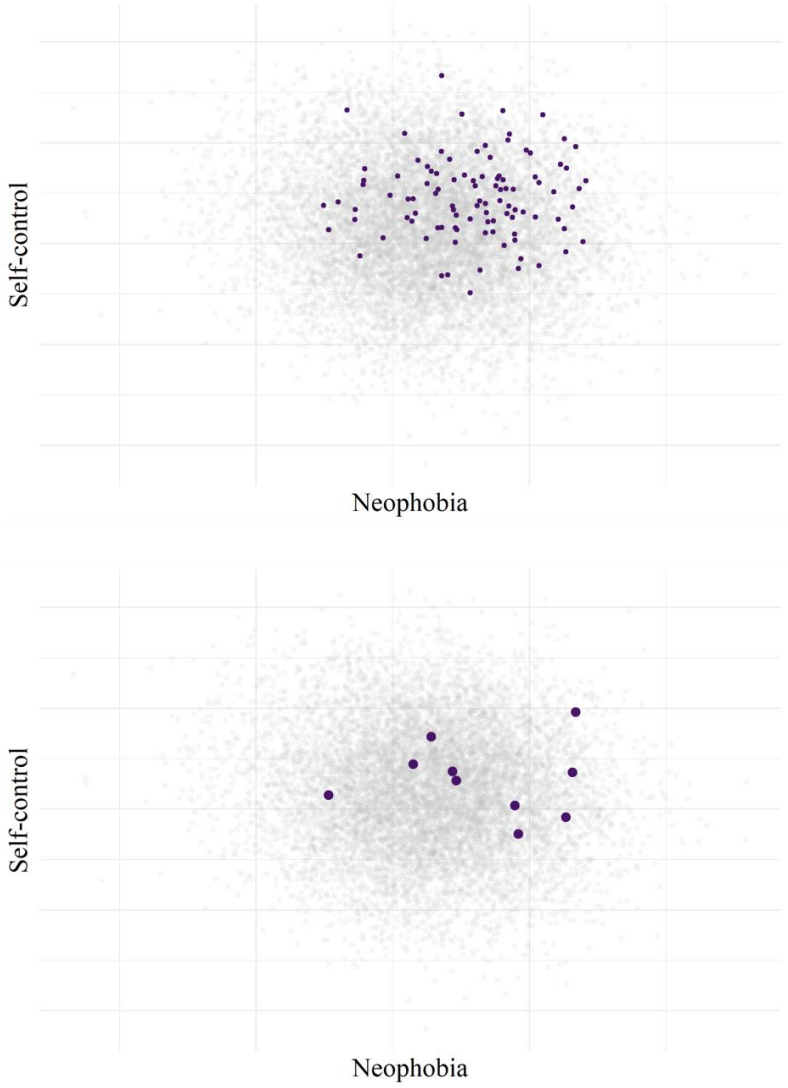


Figure 3: The Behaviour-Space of a Simulated Population of 10,000 Chimpanzees (grey dots in both panels). In purple, the Upper Panel shows 100 hypothetical chimpanzees sampled from a single site, and the Lower Panel shows just 10 of these chimpanzees.

To create a replication study, I repeated this process, taking another random sample of 10 chimpanzees from a different site. This sample is plotted in Figure 4 alongside the first sample, creating a within-species (or experimental unit) replication, which could also be framed as a between site comparison, or a test of generalisability across sites.

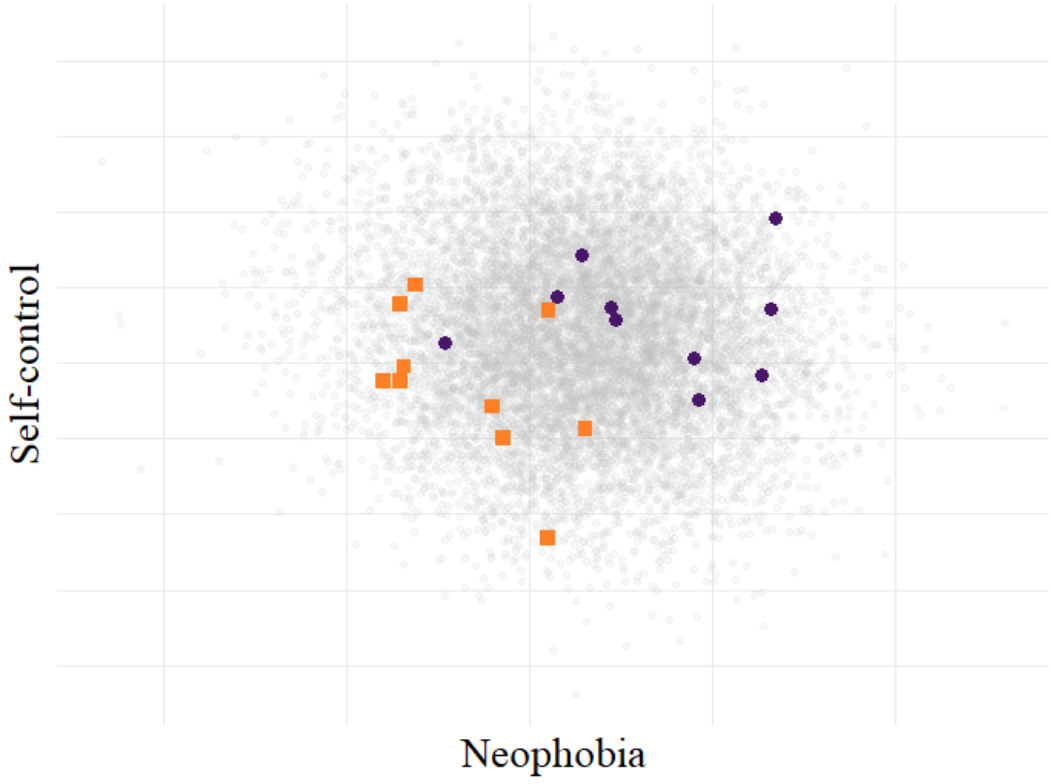


Figure 4: A Hypothetical Within-Species Replication, or Between-Site Comparison. Purple points represent the same chimpanzees sampled from the first site (Figure 3), and orange sqaures a second sample of chimpanzees.

The second sample of chimpanzees, in orange, had smaller neophobia and larger self-control scores than the first sample, in purple. Performing a two-sided Welch's *t* test, both differences were statistically significant, $p_{\text{neophobia}} = .0005$ and $p_{\text{self-control}} = 0.04$. This reflects the real variation between the sites, which were simulated at 28% for neophobia, and 14% for self-control. The samples of just 10 animals captured this difference relatively accurately, estimating the group differences as 31% for neophobia and 14% for self-control. While the two samples provided good estimates of the true between-sample differences, the samples were poorly representative of the overall population of chimpanzees. Site 1 (purple), overestimated neophobia by 14% and self-control by 3%, whereas Site 2 underestimated neophobia by 17%, and self-control by 11%.

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

$$\text{Neophobia}_{\text{bonobo}}$$

$$\beta_0 = 950$$

$$L_{0l} \sim N(0, 100)$$

$$S_{0s} \sim N(0, 100)$$

$$e_{ls} \sim N(0, 50)$$

The decision to make bonobos more neophobic than chimpanzees was arbitrary, and most empirical data supports the opposite conclusion (e.g. Forss et al., 2019). The average self-control scores were kept the same between species. Just as with the replication, I simulated all 10,000 chimpanzees and bonobos, and selected a site at random from which I sampled 10 chimpanzees, and a random site from which I sampled 10 bonobos. Figure 5 shows the results: the entire population of 10,000 chimpanzees in light blue and 10,000 bonobos in grey, and the samples are highlighted.

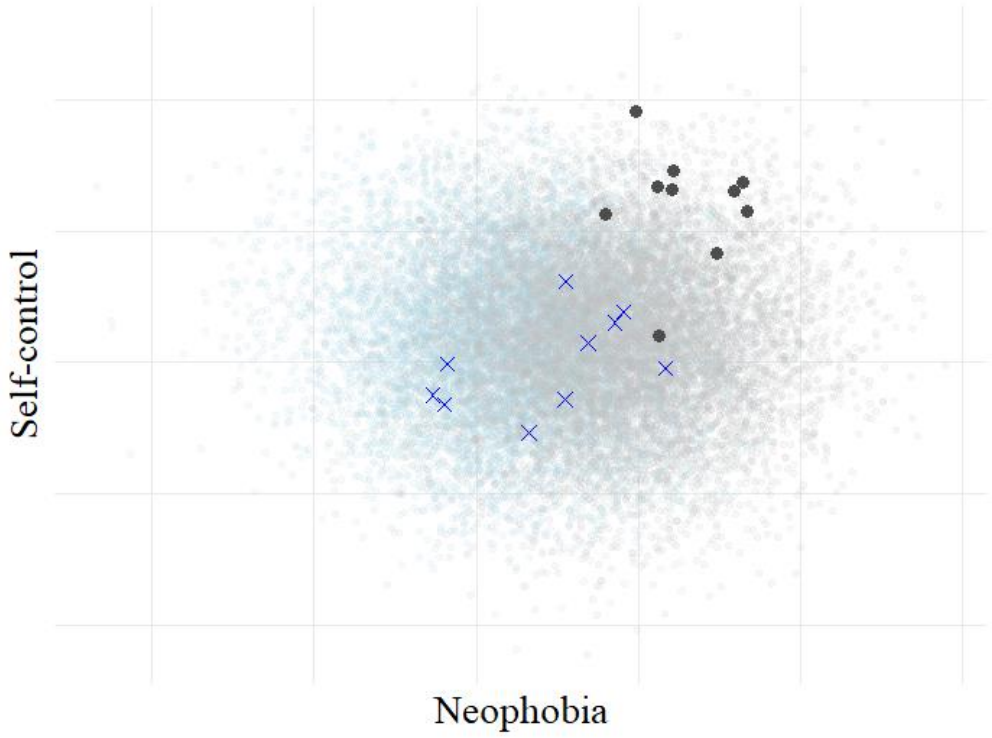


Figure 5: A Comparison Between Hypothetical Samples of Chimpanzees and Bonobos Populations of 10,000 chimpanzees (light blue) and 10,000 bonobos (grey) sampled from 100 simulated sites. Samples of 10 chimpanzees and 10 bonobos from a single site are overlaid for chimpanzees (blue) and bonobos (dark grey).

The samples in Figure 5 captured the direction of the population difference in neophobia scores, which were statistically significantly larger in the bonobo sample than the chimpanzees, $p_{\text{neophobia}} = .0004$. However, the magnitude of this effect was overestimated by 41%. For self-control, where no population differences were simulated, the samples produced a statistically significant difference between chimpanzees and bonobos ($p_{\text{self-control}} < .00001$), incorrectly estimating a species difference of over 40%. This highlights how even when a statistically significant difference is observed between species at different sites, it does not mean that the difference should be attributed to species identity alone. To explore this further, I investigated how often my comparison would return a statistically significant difference between the neophobia scores and self-control scores of the chimpanzee and bonobo samples. Because my simulation specified that there were no true differences between the species in self-control, this can provide the base-rate of false positive results, under the assumption that statistically significant results would be taken as evidence for a species difference. I simulated 100,000 comparisons between samples of 10 chimpanzees and 10 bonobos, each taken from a new site.

Across the 100,000 simulated comparisons the small sample design detected a true difference between chimpanzees and bonobos in neophobia 66% of the time with alpha = .05, which looks quite promising. However, the 100,000 simulations also detected a difference between the chimpanzees and bonobos on the self-control measure 49% of the time, in which there were no species differences specified.

Figure *6* (upper panel) plots the *p*-value distributions of the two comparisons, and the similarity between these distributions shows that observing a statistically significant difference between two samples, even if $p << .05$, is not necessarily strong evidence of an overall species difference. Figure 6 (lower panel) displays the degree of over- and under-estimation of the neophobia effect size across all samples. Strikingly, in 32% of comparisons the effect size was overestimated or underestimated by over 100%.

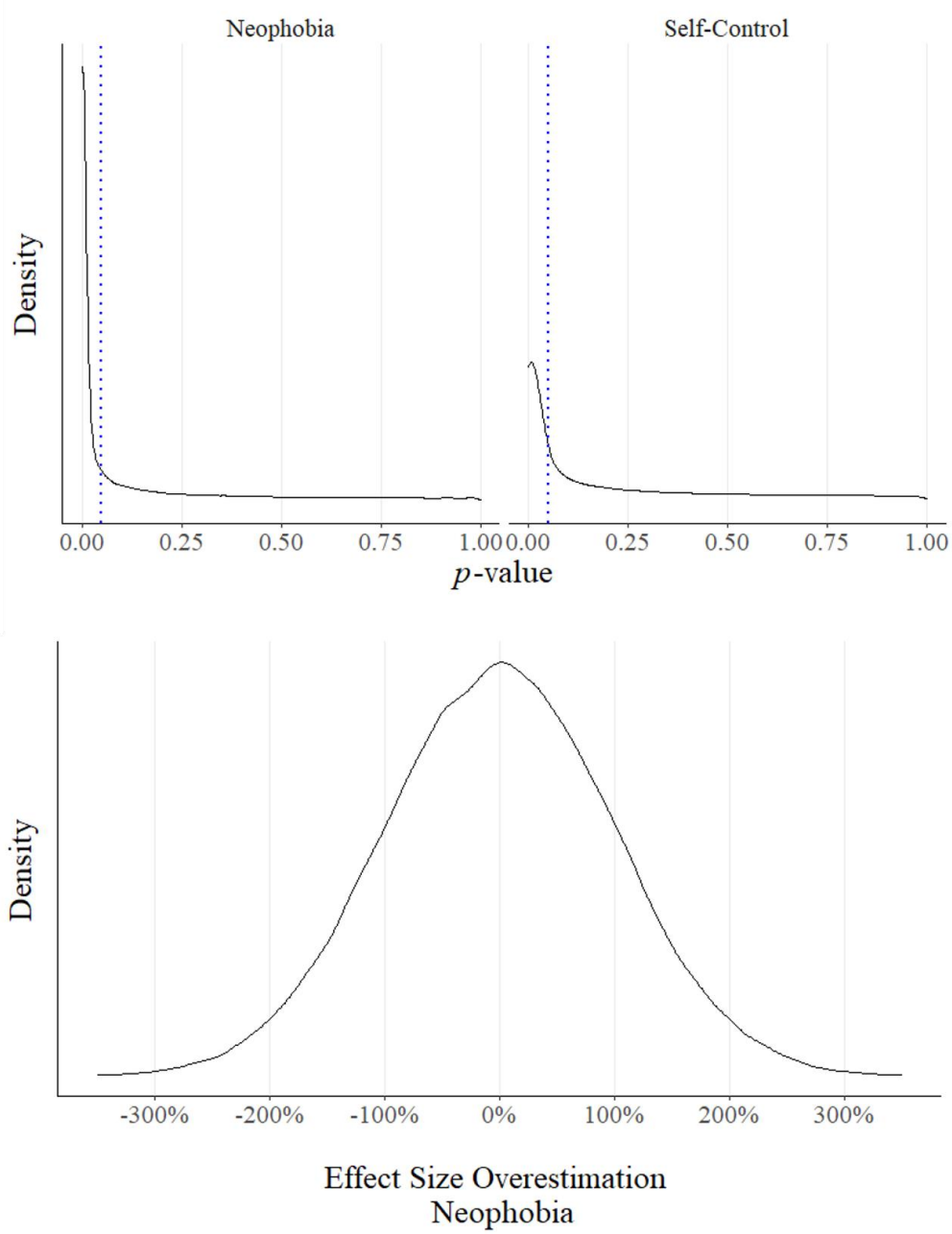


Figure 6: p-value Distributions and Effect Size Overestimation from Two Simulated Comparisons. Upper panel: *p*-value density distributions of two-sample *t*-tests from 100,000 comparisons between 10 hypothetical chimpanzees and 10 hypothetical bonobos, sampled at different sites. Lower panel: The density distribution of effect size overestimation for the 100,000 comparisons of neophobia behaviour. No data are shown for self-control as the set difference was 0, therefore it was not possible to calculate the % overestimation.

3.8. Strong and Weak Comparisons

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

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

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

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

3.9. Improving Sampling in Animal Research

There are several methods researchers can use to assess and model the effects of biased sampling on the reliability and generalisability of their research findings:

3.9.1. Experimental Design

3.9.1.1. Increasing Heterogeneity

Increasing heterogeneity is a direct method of increasing the representativeness of a sample to a target population. By sampling more diversely from within the populations specified by a theory or claim, researchers can better estimate the population parameters of interest (Milcu et al., 2018; Voelkl et al., 2018, 2020; von Kortzfleisch et al., 2020). This could involve sampling from multiple sites, such as in large collaborative studies (Crabbe et al., 1999; Culina, Adriaensen, et al., 2020; Lambert et al., 2021; Many Primates, Altschul, Beran, Bohn, Caspar, et al., 2019), but also by using multiple different experimenters and varying the conditions and treatments within sites (Baribault et al., 2018; Richter et al., 2010; Wurbel,

2002). As an example, Rössler et al., (2020) compared the ability of a sample of wild-caught Goffin's cockatoos and a sample of laboratory-housed Goffin's cockatoos to physically manipulate an apparatus to access a reward. However, rather than presenting the cockatoos with a single apparatus, they were tested in an area with a total of 20 apparatuses. Because Rössler et al. sampled from a diverse range of treatments, we can be confident that - at least for these samples of cockatoos – the results are robust across variations in treatment. An ideal experiment might generate diverse samples across all feasible factors – sites, treatments, experiments, times of day, measurements etc., which will increase the replicability and generalisability of the results (Würbel, 2000), however, it is high-cost (Davies & Gray, 2015; Mook, 1983; Schank & Koehnle, 2009).

3.9.1.2. Increasing Homogeneity and Control

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

3.10. When Does Representativeness Matter?

The resampling account highlights how for effective and reliable research, researchers should sample effectively from across their populations of interest. A representative sample is a sample with characteristics that closely resemble the target population's. Immediately, one might assume that to generate the most representative samples requires a greater number of individuals, which for many animal cognition studies might involve increasing a sample size from around 8 (Farrar et al., 2020; Many

Primates et al., 2019) to around 15 (or whatever the maximum number the researchers can access is). However, increasing the number of animals in tests is likely the least effective method of increasing representativeness in animal cognition research. This is because, i) samples at a single site are usually so unique that increasing the number tested at that site likely only increases "representativeness" by a minimal amount, but more importantly, ii) individual animals are often the biological unit of interest for animal cognition studies, and not the average response of a sample. Psychological effects happen at the level of the individual (Craig & Abramson, 2018), and robust evidence should be sought primarily at the individual level by default (Smith & Little, 2018), unless there is both, i) clear justification why group parameters (beyond checking that an individual isn't a clear outlier) have important theoretical meaning, and ii) that it is feasible to perform reliable research on these larger groups given the resource constraints animal cognition scientists face. One exception to this, covered in Chapter 2, are cases where animals cannot be exposed to multiple trials for test validity purposes, and here many animals are required to compensate for the fact that statistical power cannot be increased by trial number. Even when it might be desirable to estimate some form of group parameter, for example to generate statistical power that couldn't be achieved within individuals because of design constraints, it is important to remember that the relationship between this group (often of 'BIZARRE or 'WEIRD' animals, see Leavens et al., 2010, and Webster and Rutz, 2020) and the "target" population, if a target population can even be specified.

3.11. Barriers to effective sampling

Concerns about replicability and representativeness have surfaced often in animal behaviour and cognition research, at a variety of levels (Beach, 1950; Beran, 2012; Bitterman, 1960; Boesch, 2012, 2021; Brosnan et al., 2013; Clark et al., 2019; Dacey, 2020; Eaton et al., 2018; Janmaat, 2019; Leavens et al., 2019; Schubiger et al., 2019; Stevens, 2017; Szabó et al., 2017; van Wilgenburg & Elgar, 2013). However, it is unclear whether any real progress has been made towards understanding the prevalence and consequences of low representativeness in these fields, and I suggest that there are four main reasons why, which are theoretical, practical, motivational and educational (see also Farrar & Ostojić, 2020). These are discussed in (Farrar et al., 2021) in detail, but coalesce with larger barriers to progress in animal cognition research that I discuss throughout this thesis in Chapters 4, 8 and 10. In the following chapter, I specifically discuss how academic incentives affect each stage of the research process in animal cognition, including the sampling processes highlighted in this chapter.

4. Chapter 4: How academic incentives can affect animal cognition research[6]

Chapters 2 and 3 focused on replication in animal cognition research, making the case that areas of animal cognition research likely contains many difficult to replicate findings. I focused on the proximate causes of low replicability – publication bias, *p*-hacking, sampling variance, etc. Chapter 4 now makes an argument about the ultimate cause of this low replicability: an academic incentive structure that selects research and researchers based on their ability to produce many impactful findings. However, because of the low evidential standards in the field, many of these impactful findings are likely unreliable and have low validity. I argue that decades of research under these conditions has made it near impossible to evaluate the strength of evidence supporting many of the claims produced by the field's empirical research, because the published information is biased, and the information needed to critically assess research programmes (such as the number of unpublished findings) is usually not available. Throughout the chapter I link this argument to other debates more frequently observed in the animal cognition literature, in particular the frequent methodological debates that continue to dominate the animal cognition literature.

4.1. The academic incentive structure

While a genuine desire to understand animal minds is likely one factor driving research into animal cognition, day-to-day academic incentives likely play as important a role in determining how animal cognition research is performed. By the time most contemporary animal cognition research was being conducted (which I loosely defines as research after Premack and Woodruff's 1978 "Does the chimpanzee have a theory of mind paper) science was in the process of being heavily metricised, commercialised and increasingly competitive (for more historical accounts see: Latour, 1987; Latour & Woolgar, 1986; Lazebnik, 2018; Ravetz, 1996; Stengers, 2000; Stengers & Muecke, 2018, and for evidence see Alberts, 2013; Campbell, 2008; Edwards & Roy, 2017; Elliott, 2014; Fong & Wilhite, 2017; Lane, 2010; Seppelt et al., 2018; Simons, 2008). In order to judge which scientists were worthy of jobs and grants within and between disciplines, measures of the quality and quantity of a researcher's scientific work and proposals were required. Particularly important to hiring and grant committees appeared to be measures of the quantity and impact of scientists works, such as the H-index, number of publications and journal impact

[6] This chapter contains material published as a pre-print in Farrar, B. G., & Ostojić, L. (2019). The illusion of science in comparative cognition. *PsyArXiv*.

factor, and the quality and ambition of research proposals, assessed by peer-review. Chapman et al. (2019) provide a detailed discussion about how metrics have been used to evaluate biological scientists, and some possible consequences of this. Specifically, they outline how metrics such as the H-index and journal impact factor have been integrated explicitly into grant and job selection committees, using the example of the Brazilian National Research Council, and also cite evidence of authorship and citation manipulation by researchers and journals (Fong & Wilhite, 2017).

While explicit criteria for high impact publications have been included by job selection committees and grant requirements (Chapman et al., 2019), such explicit requirements may be becoming less visible due to initiatives such as the San Francisco Declaration on Research Assessment (DORA - https://sfdora.org/) – signers of which, including the University of Cambridge and the UKRI, pledge not to use journal-based metrics as measures of individual research article or researcher quality. However, survey studies suggest that the pressure to publish, and publish in certain journals, is still perceived by researchers. Frias-Navarro et al. (2021) asked a sample of Spanish psychologists about their perceived pressure to publish in high-impact journals and perception of competitiveness in university academic activity on a 0 (no pressure/no competition) to 10 (very strong pressure/competition) Likert-scale response. The modal response for both questions was 10, and the median was 9. These findings are in line with earlier surveys: In 2009 van Dalen & Henkens (2012) reported that that over 70% of demography researchers in the USA, UK, Canada and Australia agreed that the pressure to publish was high, and in 2019 a cross-sectional study of Dutch researchers (Haven et al., 2019) found a consistent negative attitude towards the pressure to publish across disciplines and ranks. Although formal data on publication pressure are scarce, they point to a continued perceived need to produce many impactful findings (and see Chapter 9 for my own survey of animal cognition researchers, many of whom noted similar pressures).

These pressures are exemplified by the misconduct cases discussed in Chapter 2, for example those of Diedrick Stapel and Marc Hauser, however the consequences of consistent pressures to publish extend much farther than the occasional case of clear research misconduct. Two simulation studies show the logical consequences of an incentive structure that promotes high output novel findings. First, in “The Natural Selection of Bad Science”, Smaldino and McElreath (2016) presented an evolutionary model of science. The authors simulated laboratories that varied based on their research practices, for example by having different likelihoods of investigating novel hypotheses and different false positive rates. Periodically, some laboratories were selected to “reproduce”, generating progeny that inherited the lab’s research methods, and some old laboratories “died”. The results of this simulation showed that selection

for high output promoted high false discovery rates and poorer methods. Second, Higginson & Munafò (2016) produced another evolutionary model aimed at understanding how researchers should spend their resources in order to maximise the "career value" of their publications. The optimal strategy was to perform many small-scale studies aimed at finding novel results, with between 10% and 40% statistical power to detect these effects.

Here, I explore how I think such an incentive structure influences animal cognition research, in particular research that attempts to make claims about the presence or absence of certain "higher" cognitive abilities in animals. I first make the case that the demand for continued high impact publications has created an incentive to confirm exceptional cognitive abilities in animals, and that this manifests as a theoretical bias within research labs with animals towards certain results (and equally sustained the presence of skeptics, often without access to these animals, countering these claims). These results are then readily generated by either false-positive research findings, or findings that lack validity. However, importantly, the main claim of this argument is not that all bold claims are false or invalid in animal cognition research, but that the current process of generating these claims means that it is often not possible to determine which are strong and which are not from the published literature alone.

4.2. Confirming animal intelligence and constructing clever animals

The primary research methods used in animal cognition research are confirmatory hypothesis tests (Gelman, 2014; Meehl, 1967; Rozeboom, 1960). When a researcher rejects the null hypothesis in favour of the alternative hypothesis this leads researchers to claim evidence for the alternative hypothesis and the substantive theory it was derived from. By contrast, when the null hypothesis is not rejected, these null results are often labeled as difficult to interpret and are often not published (see Chapter 8 for a study of how animal cognition researchers interpret negative results). This asymmetry allows (well-meaning) animal cognition researchers to construct hypotheses with the only real possibility of eventually confirming them. In theory, if a researcher wanted to confirm the presence of cognitive ability X in species Y, all that is required is to repeatedly test this and publish only the positive, theory-confirming results (Ioannidis, 2005; Nissen et al., 2016).

Animal cognition research may be biased further because the *direction* of confirmation, for many research programmes, is primarily focused on confirming the presence of more and more exceptional

cognitive abilities in animals.[7] This directional bias is present before data collection begins, and it can be seen already when researchers design their hypotheses (Klayman & Ha, 1987; Loehle, 1987; Nickerson, 1998; Wason, 1960, 1968). When designing a study, comparative researchers appear biased toward imagining only behaviour(s) that would be *consistent* with an animal having the target cognitive ability. Hypotheses tests are then built around this identified behaviour A, the nominated indicator of cognitive ability X, and an experiment is designed with the following null and alternative hypotheses:

H_0**:** Does not display behaviour A

H_1: Displays behaviour A

Then, if H_0 can be rejected following the experiment and analysis, researchers proceed to infer H_1 and corroborate the substantive claim and cognitive theory motivating it. Conversely, if H_0 is not rejected, generally no firm claims are made with respect to the cognitive theory and various alternative hypotheses may be discussed, ranging from less exceptional cognitive theories, to lack of statistical power and failures of the experimental method and/or auxiliary assumptions (a prominent theme that appeared in animal cognition researchers' responses to my survey, Chapter 9). There are a vast number of plausible, but untested, post-hoc reasons why a study could have "failed". For example, what if the animals were not paying attention? What if the study design was too insensitive to detect a real effect? What if the animals were not motivated to engage with the task? When these reasons are considered, it becomes clear that in many animal cognition studies, *only the positive result counts.* The exceptional cognitive theories are seldom subjected to risky or severe tests because of the distance between the theoretical claim presented and the statistical hypothesis it is tested by, and by the relative ease of explaining away "negative" results (Mayo, 2018; Meehl, 1990; Popper, 1962). In such a system, buttressed by publication bias, the eventual "confirmation" of more exceptional cognitive abilities in animals is almost inevitable.

An incentive structure that prizes publication and impact over scientific rigor, has clear negative effects throughout animal cognition research, regardless of the "direction" of the bias that manifests. However, the potential for academic incentives to bias animal cognition research is most clearly illustrated by the case of research programmes engaging in top-down drive searching for evidence that certain animals are more intelligent than previously assumed. Such a research programme is glamorous – in terms

[7] Although in the case of animal-human comparisons there appears a similar incentive, in some cases, to claim "higher" cognitive abilities in humans at earlier and earlier ages (e.g., see recent replication failures across developmental psychology, e.g., Poulin-Dubois et al., 2018 and Oostenbroek et al. 2016, and also Buckner 2013 and Bard & Leavens, 2014).

of attracting media attention and high impact publications – and can inexorably produce more and more data suggesting their animals possess certain abilities. However, the evidence they generate may not be best explained by the cognitive abilities of the animals in question, but rather are best explained by the nature of the results being set before the experiments are performed, and then the actual research acts more as a negotiation as to *how* the conclusion will be reached, not *whether* it will be reached. For example, consider you are running a hypothetical research programme that takes the following course:

1. Find a novel study species, or a novel research question
2. Find a way to reliably produce data with your species
3. Operationalize intelligence in a way that is easy to test in your species, or find existing task that you can adapt
4. Run many studies with high false positive rates and/or overinterpret true positive results; don't publish negative results
5. Make impactful claims, but add in some caveats to deflect criticism
6. "Chain" (Barrett, 2015) your species to the one above it on animal cognition's *scala naturae*
7. Fit an evolutionary account of the evolution of intelligence to your species
8. Generate funding and from this groundbreaking research, repeat steps 4-7, switching research questions in response to diminishing returns.
9. Use publication bias and the difficulty interpreting negative results to protect the core claims of the research programmes, and increase tolerance to negative results once the animal has been proven clever

Such a research programme would produce a published literature of primarily positive results in favour of certain cognitive abilities in your species and is fully compatible with known research methods and analysis practices across science from the past decades. A minimal requirement for animal cognition research programmes that make such claims should therefore be to demonstrate how the output of their research programmes are inconsistent with Steps 1-9 above. Without such evidence it is reasonable for researchers to not subscribe to the conclusions of these research programmes. For me, this proof would require: i) a retrospective re-evaluation of the evidential strength of published findings, including quantitative risk of bias assessments and systematic reviews (see Chapters 5, 6, 7, and 8); ii) replication studies of key, but statistically uncertain, findings (Field et al., 2019; Isager et al., 2021); and iii) test development, validation and triangulation using best practice and transparent methodology, including study registration.

In the next sections of this chapter, I outline why the onus should be on the researchers making bold claims about animal cognition to provide more evidence for these claims, because of the likelihood that, i) the literature contains many false positive results, ii) the current ineffectiveness of interpreting negative results, and iii) the current ineffectiveness of methodological criticism. I then argue that current scientific practices in these areas of animal cognition research generate the illusion of a scientific process, where well-meaning researchers following the textbook scientific method can consistently produce inaccurate conclusions about animal minds. I link this to discussions about the validity of tasks designed to test animal cognition, and then end by discussing the features of animal cognition research that exacerbate and sustain the current (ineffective) research practices.

4.3. False positive results

By themselves, confirmatory research methods might not lead to a greatly misleading literature, even when the direction of this confirmation is biased. However, when these confirmatory research methods are combined with a high rate of false discovery and a publication bias hiding (at least a large proportion of) negative results from the literature, then the literature can end up containing few identifiably meaningful reports. In this section, I put forward reasons as to why the rate of false discovery is likely high in animal cognition, such that the results of any given paper might be invalid.

The case that animal cognition research contains many statistically false positive findings can be made by analogy based on two overlapping observations in related fields that were highlighted in the Introduction and in Chapters 2 and 3. I briefly recap these and then present the small amount of direct evidence in animal cognition research itself.

4.3.1. Indirect evidence

4.3.1.1. Animal cognition research uses similar research and analysis methods to those used in related fields with a known high number of false positives, as shown by the results of large-scale replication studies

As highlighted in Chapter 1, large-scale replication studies across human psychology around 60% of replication studies have returned a significant result in the same direction as the original study. While there are several definitions of a successful replication (e.g. Etz & Vandekerckhove, 2016; Patil, Peng, & Leek, 2016), replication success appears low specifically in those fields that share similar properties with animal cognition, such as infant studies: small sample sizes and noisy. This replication rate of around 60% could provide an anchor for estimating replication rates of studies in animal cognition. However, as discussed earlier in the thesis, these large-scale internet/paper-based studies have limited generalisability

to animal cognition. More noteworthy for animal cognition research, however, should be the fact that non-verbal studies of infant theory-of-mind (Poulin-Dubois et al., 2018), behavioural assays of laboratory mice (Crabbe et al., 1999), and the blocking effect all have struggled to replicate (Maes et al., 2016). In lieu of evidence to the contrary, we should not expect the replicability of animal cognition research to be stronger than in these fields (after accounting for known markers of replicability such as trial number and *p*-values, see Chapter 2).

4.3.1.2. Surveys of researchers in related disciplines suggest that analytical practices that increase false positive results are used at non-negligible rates, a notion reinforced by systematic studies of analysis and reporting practices across the fields

Self-report surveys indicate that many researchers report to have used, and suspect that others use, so-called "questionable research practices", which elevate the rate at which positive evidence can be presented (Fiedler & Schwarz, 2016; Fraser et al., 2018; John et al., 2012). While some of the surveys may have biased the reported results towards higher figures (see Fiedler & Schwarz, 2016), the evidence from the various surveys shows that there is a non-trivial usage of these practices. The rates of some of these practices can also be investigated through meta-research projects, an interdisciplinary approach to evaluating methods, reporting, reproducibility and the evaluation and incentives of research (Ioannidis, 2018; Ioannidis et al., 2015). Thus far, meta-research projects and secondary data analysis projects have largely corroborated the findings of survey projects across fields (e.g., Chuard et al., 2019; Gibbs & Gibbs, 2015; Head et al., 2015; Nieuwenhuis et al., 2011; Sena et al., 2010).

4.4. Direct evidence

4.4.1. Failed replications

Because animal cognition research has not targeted systematic replication studies, little is known about the 'rate' of successful replication in animal cognition, and, as outlined in Chapter 2, such rates are likely uninformative due to the level of heterogeneity in the field, and, as outlined in Chapter 3, it is difficult to define exactly what should be considered a replication because of the theoretical openness of the field (Boyle, 2021). Nevertheless, failed replications are accruing in the field, as the examples of Amodio et al. (2021), Crosby (2019), O'Neill et al. (2021) and Soler et al. (2020) highlighted in Chapter 2 demonstrate. Perhaps most significantly, however, is that studies which we should have high a priori confidence in are sometimes struggling to replicate. Two examples are Amodio et al.'s failure to replicate Legg and Clayton (2014), and Maes et al.'s (2016) consistent failure to elicit the blocking effect with mice. Both original effects were foundational, and ones that are likely true statistical effects at the population

level, e.g., very few people would doubt the existence of the blocking effect in general. Similarly, Legg and Clayton's claim that Eurasian jays prefer to cache behind opaque objects than transparent ones when observed by potential pilferers is a logical discovery: If cache protection strategies do exist in corvids (e.g., Bugnyar & Kotrschal, 2002; Clary & Kelly, 2011; Dally et al., 2006; Emery et al., 2004; Emery & Clayton, 2001), this is one of the most likely to exist.

4.4.2. Heterogenous repeatability

One area in which measures of sampling variance have often been considered in animal cognition is the repeatability of certain behaviours within individuals, such as approaching novel objects. These studies have yielded largely consistent results: behavioural pre-dispositions, such as novel object approach, are repeatable within individuals, but not necessarily with large effect sizes. Larger studies and meta-analyses do report statistically significant, but weak, repeatability , but with considerable between study heterogeneity for behaviours such as novel object approach and discrimination and reversal learning (Cauchoix et al., 2018; Takola et al., 2021). In smaller-scale studies, this leads to non-significant repeatability estimates (e.g., Vernouillet & Kelly, 2020), and occasionally stronger evidence of near-zero repeatability of behavioural traits (such as lateralisation in detour tests in fish (Roche et al., 2020)). This low magnitude of repeatability within individuals should cap researchers' expectations of replication rates from studies using experimental designs. However, in itself, and combined with the failed replications mentioned earlier, the low repeatability of some behaviours does not mean the animal cognition literature is populated with false positive or overestimated results – just that animal behaviour is more variable than claims in the literature might imply. Next, I review the thin evidence that many of the results are likely false positive or overestimated.

4.4.3. Publication bias and low power

As discussed in Chapter 2, a combination of publication bias and low power leads to a literature of overestimated results. However, there has been little formal assessment of either in animal cognition research – although evidence of publication bias is presented in Chapters 5, 6, 7 and 9 of this thesis. Again, in the absence of direct evidence we should expect animal cognition research to reflect human psychological research, which is affected by both low power and publication bias (Button et al., 2013; Fanelli, 2012; Scheel et al., 2020), although with heterogeneity between research fields (Nord et al., 2017). Most of this heterogeneity likely lies in trial number, as animal cognition studies routinely have sample sizes of less than 10 animals (Farrar et al., 2020; Lambert et al., 2021; Many Primates, Altschul, Beran, Bohn, Caspar, et al., 2019). For the low-sample few-trials studies in animal cognition, there is a threefold

cost to low power. First, researchers risk wasting resources performing studies that will produce very uncertain data. Second, published positive results will overestimate effect sizes (Chapter 2). Third, the likelihood that positive results are a consequence of confounds in study design increases if the confounding effects, such as dominance, side biases, satiety, training effects and experimenter effects, have larger effect sizes than theoretically interesting effects, and researchers sometimes do not detect confounds in their designs.

4.4.4. Observations of probable false-positive inflating research practices

In addition to low power combined with publication bias, overestimated or false-positive findings are more likely to enter the literature if false positive inflating research practices are used. These practices are difficult to observe directly or study systematically; for example, the selective reporting of analyses is, by definition, unobservable form the paper alone. However, several researchers have documented cases where error rates may have been artificially inflated. For example, van der Vaart & Hemelrijk (2014, pp. 345–346) described how researchers in primate theory of mind research may subset their data until significant results were obtained and researchers in corvid theory of mind may select outcome variables based on their significance. Similarly, Ghirlanda (2017) noted that the statistical significance of an effect in squirrel monkey artificial grammar learning depended on the exclusion of certain data points and the weakly justified dichotomisation of an continuous variable. Further evidence of inflater false positive rates come from Waller et al., (2013) who reported that 38% of 551 primate communication studies analysed psuedoreplicated data as if they were genuine replicates. To these observations of possible error inflation, many more examples can be observed in the animal cognition literatures (I include the following examples to illustrate this not because they are likely to be false positives or especially egregious, rather they have been selected because the errors are either transparent and not severe, and where possible I have chosen examples from researchers I know personally): Researchers infer the absence of an effect based on non-significant results (see Chapter 7); analyse psuedoreplicated data as if it were independent (Tornick et al., 2016 Experiments 1 and 2; Yocom & Boysen, 2011 Experiment 1), selectively interpret the results of statistical tests (Brosnan & de Waal, 2003 cited in Wynne, 2004); drop data points to achieve statistical significance (in this case transparently in Cheke & Clayton, 2012); infer that there is a difference between two groups without performing the appropriate tests (e.g. Emery & Clayton, 2001 lack of a formal comparison between inexperienced and experienced jays; Krupenye et al., 2017 lack of a formal comparison with the new control condition and the conditions of Krupenye et al. 2016) or fail to pick up on errors in analyses that produce spurious significant results (see correction to Canteloup & Meunier,

2017). Systematic studies would be needed to estimate the exact prevalence of each of these practices or mistakes – although these are labour intensive and could likely focus on a single error only (see Chapter 7). However, I believe that collectively, the direct and indirect evidence suggests that inflated error rates are frequent in animal cognition – they are very hard to control without effective pre-registration (see e.g., Lazic, 2021 for the case of multiple comparison corrections) – and a more reasonable estimate of the false positive "rate" would be closer to .10 or even .15 than .05.

Moreover, experimenter bias further increases the possibility of false positive results in animal studies. For example, across 79 studies of aggression in ants, van Wilgenburg and Elgar (2013) reported a large difference between results when observers were reported to have been blinded to the experimental conditions or not. Converging experimental evidence was reported by Tuyttens et al., (2014), who found differences in the scoring of pig, cattle and hen behaviour depending on the conditions the coders believed the animals to be in. Compounding this, the overall rate of blinding in animal behaviour experiments appears to be low (Burghardt et al., 2012, see Chapter 6 for a longer discussion).

4.5. The inability to assess evidence of absence

While researchers might generate general evidence "supporting" certain cognitive abilities in animals at a rate above the nominal 5% false positive rate, there exists an asymmetry in which researchers rarely produce general evidence *against* the same cognitive abilities in animals. Currently, when researchers do claim that an animal likely does not have a certain cognitive trait, this is either based on the continued absence of evidence of the ability in the literature, or because the results of a single study appear to refute the presence of the ability in that species. The former method is not sufficient as the relative lack of supporting evidence in the literature is dependent on a host of information unavailable to outside researchers, such as the intensity of research and the amount of unpublished work. The latter, that a single study effectively refutes the presence of the ability in the species, is also rare. There are many plausible reasons why animals might fail to show behaviour consistent with an ability in a single experiment (Mitchell, 2014), from a failure to detect a real effect due to low power, a lack of motivation, a failure of experimental design or implementation, to the housing or rearing conditions simply not allowing the ability to manifest in those animals, even though they were theoretically capable (Boesch, 2021). Hence, researchers wanting to assess the evidence of absence of abilities in animals face an impossible task. Because of publication bias, they likely are aware of only a small proportion of the total negative-result studies, and within these negative studies they must assess the likelihood of several plausible hypotheses that are equally compatible with the published findings. This is in stark contrast to

claims of evidence supporting certain cognitive abilities in animals, which are often made from single studies with a significant result, but little test validation.

Occasionally researchers do attempt to provide evidence against the presence of more exceptional cognitive abilities in animals. For example, Suddendorf and Collier-Baker (2009) conducted a mirror-mark test in which gibbons did not use a mirror to locate hidden marks above their brow. Crucially, they also provided evidence that if the gibbons were capable of self-recognition, it would have been unlikely that the gibbons would not have used the mirror to locate the marks. They did this by conducting additional tests, demonstrating that the gibbons were strongly motivated to retrieve edible marks from their own bodies and mirror surfaces. From this, Suddendorf and Collier-Baker made the reasonable prediction that if the gibbons, in general, recognised themselves in the mirror then *most* of the gibbons would have retrieved the mark from their foreheads. In contrast, 0 of 17 gibbons retrieved the mark.

Although this inference was warranted, it was not formalised. Instead, Suddendorf and Collier-Baker's inference was supported by accepting a null hypothesis, technically a statistical fallacy. However, there is a formal analysis that could have been performed to formally reject the claim that gibbons recognise themselves in the mirror. By ensuring that the gibbons were highly motivated and capable of retrieving the mark, Suddendorf and Collier-Baker had, statistically speaking, demonstrated that only large effect sizes were consistent with the hypothesis that gibbons recognise themselves in the mirror. Equivalence tests could then be used to reject the presence of effect sizes of theoretical interest from this study, a topic discussed in Chapter 7's exploration of how animal cognition researchers interpret negative results.

4.6. Interim Summary: The Illusion of Science and Methodological Criticism

Thus far, I have argued that many literatures claiming to support certain cognitive abilities in animals are compatible with research programmes generating many false positive results, or true positives with little validity. Such research practices have been promoted by an academic incentive structure selecting research that produces many, novel and positive findings over scientific rigour (Higginson & Munafò, 2016; Smaldino & McElreath, 2016). Of course, the literatures are also compatible with the animals having the cognitive abilities in question, too – the message of this chapter is that currently the research required to distinguish between these two possibilities, replication studies and systematic meta-studies and meta-analyses, are seldom being conducted, and when they are conducted, they are invariably difficult to interpret (with some exceptions in some subfields, see Chapter 8). If, the argument in this chapter is correct, the question becomes what has prevented greater explicit recognition of the directional biases and confirmatory research practices across areas of animal cognition? In my view, this is because the

current scientific training and processes create the illusion of a scientific process that will inevitably produce cumulatively stronger findings and converge on the "truth" (Bloor, 1974), without this necessarily being the case: Researchers generate hypotheses, operationalise variables, construct detailed study plans, perform statistical analyses, submit findings to respected journals and undergo peer review, and embrace a culture of methodological and conceptual criticism. But none of these stages are necessary or sufficient for scientific progress, and biases can accumulate across each stage (Meehl, 1990; Munafò et al., 2017). Researchers are often blind to these biases through their own training, often statistics and using null hypothesis significance testing (Gelman, 2014; Gigerenzer, 1998a, 1998b, 2004; Gigerenzer et al., 2004; Lambdin, 2012; McShane et al., 2019; McShane & Gal, 2016), but also in how science works (Forrt, 2019; Gelman, 2014; Koroshetz et al., 2020; Stengers & Muecke, 2018). These training biases are compounded by a host of psychological biases that further blind researchers to problematic research practices: confirmation bias (Nickerson, 1998), survivorship bias (Smaldino & McElreath, 2016), pluralistic ignorance and social norms, the sunk-cost fallacy, motivated ignorance, motivated reasoning, status quo bias, status and authority bias (see Jussim et al., 2019 for an overview).

However, one may object to this line of argument and argue that animal cognition research effectively elicits criticism of its methods and findings. Criticism and nuances of operationalisations and task designs are omnipresent in the literature (e.g., Anderson & Gallup, 2015; Heyes, 2015; Lind, 2018; Povinelli, 2020; Povinelli & Vonk, 2004; Redshaw et al., 2017; Suddendorf & Corballis, 2008; Vonk, 2019), and perhaps this strong methodological criticism combats the issues of confirmatory research practices and directional bias. However, when this criticism is, a) primarily methodological, and b) focused on the individual study, the criticism can become ineffective in the long-term, even if it is scientifically valid, because it can promote incremental changes to a flawed process. That is, it promotes the illusion of a cumulative scientific process, where researchers respond to criticism iteratively with "improved" methods, where in reality the same errors of directional confirmation are repeated. Rather than being a feature of a rigorous scientific process, the prevalence of so much methodological criticism can be viewed as a symptom of a poorly functioning scientific system, one that focuses superficially on task design over a greater focus on test development, theory, measurement and construct validity (Allen, 2014; Borsboom, 2006, 2014; Eronen & Bringmann, 2021; Farrell & Lewandowsky, 2010; Flake & Fried, 2019; Michell, 1997). A full

discussion of these issues in animal cognition research is overdue and beyond the scope of this thesis but is touched upon again briefly in the discussion of modelling in Chapter 10.

4.7. Not just clever animals, not just top-down research

The argument in this chapter has been centered around research programmes that have inexorably produced evidence "supporting" the presence of complex cognitive abilities in animals, which often use a top-down approach that has been criticised elsewhere, too (de Waal & Ferrari, 2010; Eaton et al., 2018; Vonk, 2021). However, much of the argument can be applied to research not aiming to prove complex abilities in animals too, whether this be over-emphasizing human exceptionalism or confirming the presence of cognitive abilities in infants at earlier and earlier stages. Critics and skeptics are under the same academic incentives to produce many impactful papers too, and this can result in oversold criticism, too. In many ways, the economy of animal cognition rests on a symbiosis between the "killjoys" and "romantics" (Shettleworth, 2010; Starzak & Gray, 2021), in which each area necessary for the others academic survival. It is also important to distinguish between incentives that act at the level of the research programme, and those at the level of the individual study. While research programmes might still be geared towards certain findings in order to appear attractive to funders (Lilienfeld, 2017), the recent shift in academic incentives away from sensationalism, such as an increase in the number of venues accepting or soliciting negative results, might promote biases in the opposite direction, e.g., overinterpretation of negative findings (Aczel et al., 2018). As the pressure to publish is still felt by researchers and barriers to publication are being reduced (e.g., through accepting negative results, replication studies, and journals such as the Frontiers and MDPI groups with "collaborative" peer review), there is a danger of more low-quality research being published, irrespective of the outcome. While removing publication barriers is important for evidence synthesis because there are many ways for false negatives to occur, researchers must be wary of seeking and overinterpreting negative results (Aczel et al., 2018).

4.8. Key Sustaining Features: Ambiguity and the Small Structure of Animal Cognition

Two key features of animal cognition research help to sustain the current pattern of research criticised in this Chapter. These are ambiguity and the small and siloed structure of the field, which I now discuss in turn.

4.8.1. Ambiguity

"It is impossible to say that technical literature always errs on the side of caution; it also errs on the side of audacity; or rather it does not err, it zigzags through obstacles" (Latour, 2003, pg. 55)

When researchers claim support for a hypothesis based on a statistical test, they implicitly claim that the statistical hypothesis is a close and valid test of the substantive theory, too (Duhem, 1976; Meehl, 1990). However, in animal cognition there is usually a large distance between the statistical and substantive hypotheses – the data frequently underdetermine theory (Boyle, 2021). It is this distance that is the locus of most methodological criticism in the field. Insofar as animal cognition attempts to test for human-like cognitive capacities in animals (Heyes, 2019), this distance can always be found (Hempel, 1958; Hennefield et al., 2018; Povinelli & Henley, 2020).

Part of the issue stems from the inherent ambiguity in the verbal hypotheses and definitions that are derived from theory.[8] Across psychology, such verbal models are open to many available and justifiable interpretations (Smaldino, 2016, 2017b), and the disagreements that arise from this are common in animal cognition (see Hampton, 2019; Hennefield, Hwang, & Povinelli, 2019; Heyes, 2019). The flexibility afforded by underspecified definitions permits results that 'cognition optimists' would perceive as novel and groundbreaking, whilst 'cognition pessimists' can equally well use the same results to invalidate the test and/or discuss the lack of evidence for the exceptional theory.

Perhaps more paradoxically, definitional ambiguity also allows the *same* researchers to flexibly adjust their substantive claims depending on whether they are refuting criticism or selling the results. As this topic is an underexplored area that is open to interpretation, I will use examples from research by the lab I am a member of. First, as Penn and Povinelli (2007) highlighted in their critique, Dally, Emery and Clayton (2006) "acknowledge, that scrub jays' ability to keep track of which competitors have observed which cache sites 'need not require a humanlike 'theory of mind' in terms of unobservable mental states". However, their title, "Food-caching western scrub-jays keep track of who was watching when", lends itself to more exceptional cognitive interpretations for those who want it, and it is these more exceptional claims that forms the basis of later claims of the potential for higher cognitive abilities in corvids (Clayton, Dally, & Emery, 2007). Another example comes from Ostojić, Shaw, Cheke, & Clayton (2013). While the

[8] Although the verbal vs formal/mathematical model debate in the psychological reform movement oversimplifies the issue. The same concerns, to different degrees, can be raised at any science using ordinary language, which is all scientific disciplines to an extent. Many thanks to Marta Halina for this point. Psychologists should be wary that just because an idea or verbal model has been translated into a more formal model, these models may have some of the ambiguity of their verbal ancestors built-in (although the process of formalising the model might make this more explicit). The utility of more formal models in everyday comparative cognition research is an interesting area for future research.

authors of this study do believe that the results of this study present a “crucial first step in demonstrating state-attribution” (p. 4127) and mean that “Eurasian jays’ food-sharing behaviour represents a useful paradigm within which to investigate whether these birds, and more generally nonhuman animals, might be capable of desire-attribution” (p. 4127 and Ostojić, personal communication), the claims of the study do lend themselves to be reported as evidence for state-attribution (e.g. Keefner 2016).

When more formal analyses have been performed, conceptual flexibility or ambiguity has often been foregrounded. For example, in research into animal gestures Bourjade et al. (2020, p. 821) concluded that “the concept of gesture suffers from several conceptual weaknesses that are; (i) various degrees of semantic ambiguity, (ii) several unacknowledged assumptions, and (iii) inappropriate classifications of what a gesture is and is not, from one study to another”. Similarly, when Colbourne et al. (2021) applied a more neurocognitively appropriate definition of tooling (Fragaszy & Mangalam, 2018) to the tool use literature – an area Colbourne et al. highlighted the conceptual looseness in – the number of reports that qualified as tooling lay much smaller than the number of claims of “tool use” in the literature (also see Bastos et al. (2021) for a discussion on less ambiguous criteria for testing tooling).

Finally, it is notable that many different claims are implicitly or explicitly made in a single paper, as was highlighted in Chapter 3’s analysis of the different claims that could be extracted and can be extracted from a single paper, in Almeling et al. (2016):

5) Socially living Barbary macaques lose interest in the non-social environment with age
6) Barbary macaques lose interest in the non-social environment with age
7) Socially living monkeys lose interest in the non-social environment with age
8) Monkeys lose interest in the non-social environment with age

While ambiguity has been criticised repeatedly in science’s reform movement (Farrell & Lewandowsky, 2010; Guest & Martin, 2020; Smaldino, 2016, 2017b; Tunç et al., 2021), an alternative perspective is that the caveats in papers, rather than being tools to deflect criticism, reflect genuine uncertainty on the part of the researchers, something to be expected in a discipline like animal cognition (Boyle, 2021). The caveating present in the body of an article might then be lost in the abstracts and titles due to journal requirements on word counts, and subsequently lost in future citations and attempts at synthesis. This can be exacerbated by memory constraints, i.e., readers (and the authors) might not remember the caveats in all of the articles they have read, and further exacerbated by the lack of methods

to synthesise very uncertain data effectively, especially when this uncertainty is only conveyed verbally in articles.

4.8.2.Small Structure

As stated earlier, many of the issues discussed in this paper are not exclusive to animal cognition and have in some form or other been discussed by different researchers in different fields. However, this chapter argues that the extent of these issues appears to be very severe in animal cognition. One of the reasons for this is the field's small size – in which often only a few researchers or a single group study a particular question in a particular species or family. In 1950 Beach lamented the limited focus of comparative psychology on learning mechanisms in the lab rat, suggesting that the lack of more comparative work (more species, and more questions) limited the progress the field was making. Now the situation is markedly different; animal cognition is no-longer over-specialised *sensu* Beach and boasts a wide range of research questions across many different species (Beran et al., 2014; Shettleworth, 2009; Vonk, 2016). However, Beach (p. 120) caveated the call for more species and more questions, highlighting that when many researchers and groups work on the same problems it becomes "*possible to check the accuracy of the findings, to accelerate the acquisition of new data, and to formulate more valid and general conclusions than could have been derived if each worker dealt with a different species*."

It is clear that Beach's warning rings true in many areas of animal cognition research today. Although not formally quantified (but see Chapter 7 for corvid social cognition), the research field is heavily siloed for many species-topic combinations, with little truly independent replication (Lambert et al., 2021; Many Primates et al., 2019; Chapter 7). This is compounded by the inevitability that "independent" groups will end up having shared group members – for example PhD students in one lab who then take up positions in others. Even when groups are "independent", they might often share a common incentive (e.g., proving that group X have cognitive ability Y) that reduces this supposed independence. Of course, the field is heterogeneous in this manner, with easy to access or easy to house animals likely having more independent groups studying them, such as dogs (Aria et al., 2021) or pigeons (Lambert et al., 2021), and fields with historical interest that have attracted more funding, too – such as primate social cognition (Halina, 2021).

When independent criticism does occur for the siloed research fields, this often comes from those not directly involved in data collection on that species. In these cases, there is a clear asymmetry between those who have access to the species in question and those who can criticise it. The critics can outline weaknesses in current data, and outline what convincing data would look like, but for many species, they

cannot collect it themselves. The groups that control the resources that are necessary to ask questions also control the answers that can be made. While skeptics require consent and collaboration to severely test claims, researchers can readily produce theory-confirming data without the consent of or collaboration with a skeptic. This is analogous to how tobacco companies could refute the link between smoking and cancer, not by performing clearly flawed studies, but by selecting which studies to fund and how to disseminate the answers (O'Connor & Weatherall, 2020).

The final issue with the small nature of the field is that the number of questions that can be asked far exceeds the amount of resources that researchers have to answer them. In response to diminishing returns in one area, for example due to exhausting easy-to-perform experiments, experimental resources can be refocused onto a new topic, which can be exacerbated by the relatively short turn over times of PhD students and employees on short-term contracts. This might be a cost-effective way of performing science if the ratio of true positive results to false positives was high, but if, as argued throughout this thesis, research methods that readily produce false positive results are being used, the field can create the illusion of productivity as lots of "discoveries" are published, without actually verifying these findings. In other words, when a research programme begins to degenerate (Lakatos, 1970), researchers are incentivised not to find the sources of this degeneration and self-correct (Ioannidis, 2012a; Rohrer et al., 2018; Vazire & Holcombe, 2021), but simply to switch research questions. This may be in contrast to more theory-driven veins of animal cognition research (e.g., Ghirlanda et al., 2017 in sequence learning) or scientific research (e.g., the ATLAS Collaboration 2021 in the search for supersymmetry in high energy particle physics), where strong predictions and models make it clear when certain predictions are being tested and when they are not (but see Smith et al. (2012) for an alternative perspective on the use of associative modelling).

In summary, when a research field is small and siloed, reliability and validity issues can be exemplified – especially due to the limited potential for independent replication. Clearly, a balancing act is required to account for both the difficulties of performing strong science with limited resources, whilst also maintaining the diversity that Beach and others pined for during the dominance of the lab rat (Beach, 1950; Bitterman, 1960). A first step would be to complete demographic surveys of the field, identifying the areas of animal cognition research that are most siloed and those with more independent research, which can then feed into risk-of-bias assessments for specific research programmes. Chapter 7 presents an early attempt of this for the field of corvid social cognition.

4.9. Summary and Pushing Back

In this chapter, I have argued that the academic incentive structure promotes misleading research in animal cognition, specifically focusing on the case of top-down research aimed at providing evidence for certain cognitive abilities in certain species - whilst noting that much of the critique applies throughout animal cognition research as a whole. The chapter claimed that, on the whole, animal cognition uses confirmatory research methods that are biased towards confirming the presence of more exceptional cognitive abilities in animals. When the directionally biased confirmatory methods are combined with false positive inflating research practices and publication bias, the literature can become filled with misleading results, and this is likely the situation in much animal cognition research. There are two-counter arguments to these positions that I now address. The first is that there is a lack of evidence demonstrating that animal cognition research does produce false positive results and as such we should not question the literature until this evidence emerges – which I counter. The second is that animal cognition research is simply too heterogenous for arguments like those made in this chapter to be easily interpreted and deployed – which I accept.

4.9.1.On the lack of evidence

The first objection to the arguments presented in this chapter so far is that little formal and direct evidence has been provided in animal cognition research that, i) there is a publication bias, ii) rates of false positive results are elevated and iii) confirmatory research methods are often used. In the absence of this direct evidence, should we continue with the *status quo* until studies have sought to evaluate these practices? Personally, I think not: the indirect evidence (through analogy with other fields) and knowledge gained through informal discussions are overwhelming, such that the default position should be that animal cognition research is equally or more affected (due to the small size issues) to the nearest other discipline that has investigated these issues (likely some combination of human psychology and biomedical fields). In other words, the onus should be on those maintaining the claims in the published literature to provide evidence that their claims are *not* the product of publication bias and false positives, rather than on the skeptic to demonstrate that they are. This can be achieved prospectively through mass transparency and “best practice” methods in future studies, including pre-registration and publishing null findings, and retrospectively through systematic review projects, meta-analyses and risk of bias assessments. However, such research projects are labour intensive, and it is unlikely that a systematic review and risk of bias assessment will be conducted for each vein of animal cognition research a person is interested in. In lieu of these projects, scientists must look to the indirect signs of reliability and validity

they can find (see Chapter 2), but in the absence of finding positive evidence about the reliability and/or validity of a finding, there is no requirement to believe it.

4.9.2.On heterogeneity

A more convincing counterargument to this chapter is to point to the mass heterogeneity of research methods employed across animal cognition, even in research programmes that produce evidence in favour of certain cognitive abilities in animals. And some areas of animal cognition research *do* have the hallmarks of strong replicability and validity (see Chapter 2). While questionable research practices may be used and the rate of replication may be low *on average,* this does not mean that any given study itself contains misleading evidence. The rate of questionable research practices likely varies not just between individuals, laboratories, research areas, but also within them, and this means that identifying areas of problematic research can become incredibly difficult. Just because a certain study on cognitive ability X in species Y was *p*-hacked, this does not mean that all other studies of X in species Y were, too. Even if a meta-analysis finds convincing evidence of publication bias within a certain area of study, this does not automatically mean that all individual studies in that area are biased. However, identifying the studies that are at low risk of bias can be difficult, if not impossible, for an outsider from the published literature alone. Again, the same prospective and retrospective strategies are needed to tackle the issue of heterogeneity, "best-practice" studies with full transparency, and systematic review projects, meta-analyses and risk of bias assessments. Chapters 5, 6, 7 and 8 focus on this latter issue. Chapter 5 presents a first attempt at assessing publication bias and statistical inferences in the animal physical cognition literature. Chapter 6 presents a systematic review and quantitative risk-of-bias assessment across nearly the entire corvid social cognition literature, and Chapter 7 specifically focuses on the interpretation of non-significant statistical results across animal cognition research. Chapter 8 acts as a collective discussion for these three chapters on the lessons I learned conducting secondary data analyses, evidence synthesis and bias detection across difference elements of the animal cognition literatures.

5. Chapter 5: Publication bias and statistical inference in animal physical cognition research[9]

This chapter marks the start of three empirical chapters aiming to describe, synthesize and assess risk of bias in evidence in animal cognition research. Because of the heterogeneity of the discipline, each chapter uses a different sampling strategy to target different elements of the literature. Chapter 5 samples a small section of a diverse literature on animal physical cognition, Chapter 6 aimed to sample all interventional studies in a narrower field, corvid social cognition, and Chapter 7 aimed to sample a specific occurrence (non-significant results from null hypothesis significance tests), from across 20 different journals publishing animal cognition research. As a first attempt at bias detection, this chapter aimed to provide a preliminary assessment of statistical design and inference, publication bias and reliability in a sample of animal physical cognition experiments, each issues that have been foregrounded in the replication crisis and reform movement of psychology (e.g., statistical inferences: Aczel et al., 2018; McShane et al., 2019; Smith & Little, 2018; publication bias: Fanelli, 2012; Scheel et al., 2020; replication: Open Science Collaboration, 2015; Zwaan et al., 2018).

5.1. Statistical Design and Inference

5.1.1. Sample Size and Statistical Biological Unit of Interest

My first aim was to evaluate the statistical design of the experiments in our sample of animal physical cognition research, specifically the sample size and the biological unit of interest. Small samples can lead to two problems. First, small sample sizes are often cited as a cause of statistical designs with low statistical power (e.g., Button et al., 2013), leading to overestimated effect sizes that are difficult to replicate and a distortion of the truth concerning the literature (Cumming, 2008; Fiedler & Prager, 2018; Hedges, 1984; Chapter 2). Second, small samples may poorly represent the researcher's overall target population, leading to over-generalized claims (Henrich et al., 2010; Hurlbert, 1984; Chapter 3). However, research with small sample sizes has been used effectively throughout the history of comparative psychology. When experiments use many trials within each individual animal, their statistical tests can achieve high power to detect theoretically interesting effect sizes. This often occurs when researchers focus on the individual animal as the biological unit of interest, rather than the group (Smith & Little, 2018). In animal cognition research, the individual may be the most meaningful unit of analysis (Craig &

[9] This chapter contains material published in Farrar, B. G., Altschul, D. M., Fischer, J., van der Mescht, J., Placì, S., Troisi, C. A., ... Clayton, N. S., & Ostojić, L. (2020). Trialling meta-research in comparative cognition: Claims and statistical inference in animal physical cognition. Animal Behavior and Cognition, 7(3), 419.

Abramson, 2018): if an animal understands properties of the physical world, effects will manifest within this individual, but not necessarily at the level of the group. For example, when learning the trap tube task in Povinelli's "*Folk Physics for Apes*", Megan the chimpanzee was correct on 80/100 trials, (p = .000000000135), no other chimpanzee performed significantly above chance (Povinelli, 2000). As a result of these inter-individual differences, it would have made little sense to focus exclusively on the group. Although the group response is interesting, the most informative analyses will also examine effects at the level of the individual (for example by performing tests within each animal, or by building models with both group and individual effects). The first aim of the project was therefore to characterize the biological unit of interest (individual or group, or both) of the statistical analyses in each paper and the sample size that the researchers tested. This approach provided a basic description of the statistical designs used in our sample.

5.1.2.Publication Bias and Statistical Reliability

My second aim was to collect data on three indicators of statistical reliability and publication bias, namely, i) the prevalence of positive claims, which can be an indirect measure of publication bias, ii) the distribution of reported *p*-values, which can give clues to the overall strength of evidence that researchers are generating against null hypotheses, and, iii) the proportion of animals "passing" any given test, which can indicate the robustness of statistical conclusions across individuals. These three measures provide data about the reliability of research findings in physical cognition; if the literature has a large publication bias, or contain many just-significant *p*-values (i.e., around the alpha = .05 threshold), it will likely hold that these findings are difficult to replicate. Similarly, when only a small proportion of animals pass any given test of physical cognition, replication studies may "fail" because they miss these individuals in their samples.

5.1.3. Prevalence of Positive Claims

Publication bias leads to a literature filled with overestimated effect sizes (Hedges, 1984; Sterling, 1959), and facilitates the canonization of false facts (Nissen et al., 2016). To assess the severity of publication bias across scientific fields, researchers have investigated whether the literature contains an "excess" of positive results. This assumes that if a literature is filled exclusively with positive results, there is likely a number of unpublished negative findings, too. In one such study, Fanelli (2010) reported that just over 90% of a sample of psychology and psychiatry papers that contained the phrase "test* the hypothes*" reported support for the hypothesis under investigation, suggesting that many studies that did not support certain hypotheses were unpublished. However, the presence of publication bias in

comparative psychology has largely been unstudied, which I aimed to examine in animal physical cognition research. However, the most often employed tests for publication bias seemed unsuitable for work in the physical cognition literature. Studies such as Fanelli's are subject to strong sampling biases (restricting analyses to articles that contain the phrase "test* the hypothes*, of which there may also be too few in smaller literatures), or tests such as asymmetry tests of funnel plots in meta-analyses. This would require identifying a body of research that it makes sense to group together, for example testing the same hypothesis. Because of the heterogeneity of animal physical cognition research (with many different species, task designs and questions being used), this was deemed inappropriate. Hence, two strategies were used. The first, I developed a method that focused on researchers conclusions – whether they claimed their experiment demonstrated some physical ability in animals or not, and the second used *p*-value distributions.

5.1.4. P-value Distributions and Evidential Strength

p-value distributions offer a window into the statistical reliability of a research field. First, if a set of studies have power to detect a predicted effect size, which turns out to be approximately true, the *p*-value distribution from across these studies will be right-skewed, i.e., there will be more *p*-values in the interval 0 to .01, than in the interval .01 to .02, and more in this interval than between .02 and .03, and so forth. In contrast, the *p*-value distribution of a body of research examining false effects will be uniformly distributed, for most tests (Simonsohn et al., 2014). Comparing the shape of a research body's *p*-value distribution to the shape of *p*-value distributions expected under different conditions therefore provides some information about the strength of evidence against null hypotheses (Lakens, 2017a; Simonsohn et al., 2014). Second, *p*-value distributions can offer a perspective on publication bias and false-positive inflating analysis practices. If the number of *p*-values reported just below .05 is disproportionally higher in the literature than the number of *p*-values reported just above .05, this suggests that either some *p*-values have been coerced into falling below the significance threshold, or that effects above the significance threshold have not been published. I therefore recorded the *p*-values supporting the main claims in our sample of physical cognition research to provide data on the strength of evidence against null hypotheses across this body of research.

5.1.5. Proportion of Animals "Passing" a Test

Finally, I coded the number of animals reported to have "passed" each test of physical cognition, if such a test was performed. This number provides information on how generalizable certain individual-

level effects are within the original sample, which may help to calibrate researchers' expectations about the likelihood of replicating effects in new samples.

5.2. Method

The working introduction and methods for this paper were deposited before data collection at https://osf.io/3d9vh, and the final dataset and coding materials are available at https://osf.io/wkpeq/.

5.2.1.Paper Inclusion

This project attempted to code information from 200 published experiments in animal physical cognition studies. This sample size was decided based on a subjective cost-benefit analysis – extracting information from enough experiments to get a good coverage of animal physical cognition research, whilst recognising the study was exploratory and aimed develop novel methods. Papers were identified using a keyword search in Scopus. Papers with titles, abstracts and/or keywords containing the following keywords: "folk physics" OR "physical cognition" were searched, returning 167 results on 26th November 2019. An error in an earlier search meant we had expected over 600 results from this search, and after realising that the search returned only 167 results, I performed a further search for "trap-tube" OR "trap tube" OR "trap table" OR "trap-table", which returned a further 58 results. Papers were listed by 'relevance,' and the titles and abstracts of each paper were then screened for whether they fit my inclusion criterion: being a study of physical cognition in captive animals, which also included animals kept in captivity transiently for testing. For multi-experiment papers and papers with many different conditions, two experiments or conditions were randomly selected for coding, using the function 'sample' in R 3.6.3. This procedure was decided to give greater representation to common designs used in animal physical cognition studies, while minimising the extent my analysis would be biased by overweighting certain studies.

Of the 167 papers from the first search, 60 were coded as fitting the inclusion criterion. An extra 12 non-duplicate experiments from the trap-tube/trap-table search fitted the criterion and were added to the sample. A second screen of the full papers then led to 9 more studies being excluded: 7 for containing experiments outside of physical cognition or being developmental or personality studies, 1 for inability to access the paper, and 1 for the paper not being written in English. Of the remaining 63 papers, 53 involved multiple physical cognition experiments, from which I randomly selected two experiments to be coded. This produced a total sample of 116 experiments from 63 papers. The 63 papers were published between 1994 and 2019, across 19 different journals. The journals with the largest number of papers in

my sample were Animal Cognition (20), the Journal of Comparative Psychology (10), and Animal Behaviour (7). A complete reference list of the papers and journals included in this project can be found in the References. There were 45 different species represented across the 63 physical cognition papers that we coded (Figure 7), with chimpanzees (8), New Caledonian crows (8), dogs (7), orangutans (6) and keas (6) being the most common.

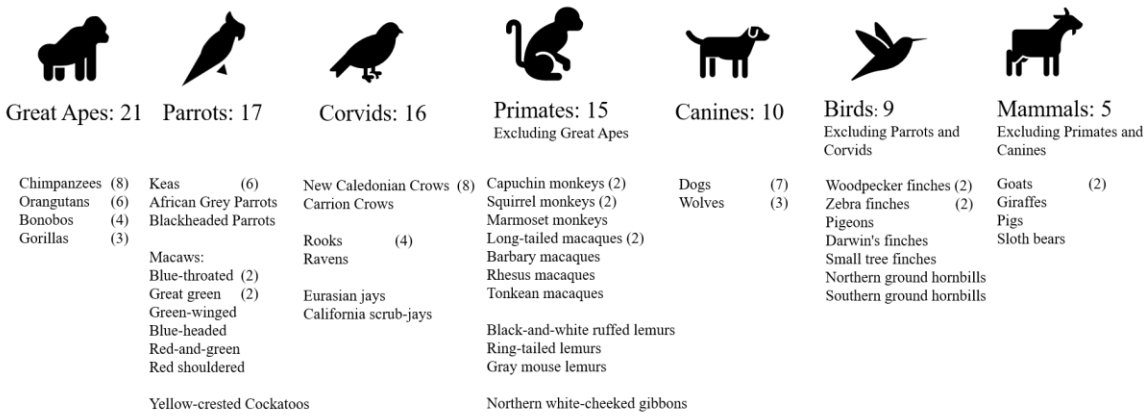


Figure 7: The number of papers investigating each species in our sample of physical cognition papers, e.g., there were 8 papers that included chimpanzees in the sample, and two papers including goats. If a species was in more than one paper, this number is given in brackets after the species.

The 63 papers contained different physical cognition tasks. Specifically: 14 papers used variants of the trap-tube task; 13 examined other forms of tool use, for example testing the ability to select and use tools based on their properties or functionality; 11 examined animals' responses in means-end, contact and string pulling tasks; 9 papers examined animals' understanding about hidden objects, for example using object permanence, object-tracking and object choice tasks; 5 papers tested animals abilities to gain access to food; 3 papers used the physical tests from the Primate Cognition Test Battery; 3 papers used support problems; 2 papers examined how birds choose nest building material; 1 paper tested dogs' understanding of solidity; 1 paper investigated tool manufacturing in crows; and 1 paper investigated both means-end understanding and tool use in lemurs.

Importantly, this sample is weakly representative of physical cognition research as a whole, and more closely aligned with research identified as "folk physics" - focusing on how animals manipulate objects to gain access to rewards, whether they respond to features of the environment such as connectivity and gravity or distinguish between functional and non-functional tools. Because of my search criteria, trap-tube and tool use studies are over-represented in my sample, some task formats are only

represented once (e.g., water raising tasks), and some are not represented (e.g., violation of expectation tasks). The sample does not include experiments on spatial cognition, or numerical cognition.

5.2.2.Coding Protocol

Two individuals coded each paper according to the protocol detailed below. An early version of the protocol was piloted on the first experiments of Povinelli's "Folk Physics for Apes" (no experiments from Povinelli's book were included in the analysis), and a second protocol was piloted on four non-physical cognition studies. For each paper, the following features were coded:

1. The main claim of the paper, coded from the abstract of each paper

Coders were asked to copy and paste the sentence(s), from the abstract, containing the main claim the authors made from each paper.

2. Whether this claim was "positive", "negative" or "inconclusive"

Coders were asked to decide whether each claim was "positive", "negative" or "inconclusive". I defined positive claims as asserting the presence of a more exceptional ability in the animal, a novel effect, or the animal "passing" a test, negative claims as asserting the absence of a more exceptional ability in the animal, or the animal "failing" a test, and all other claims were labelled inconclusive. In addition to these definitions, coders were provided with more information about what would constitute a positive claim, namely "whether the paper was claiming that the animals are 'clever' or passed some criterion of physical cognition." Alternatively, a positive claim could be a negative result in a control condition e.g., "the animal's performance could not be explained only by a simple rule"), and an inconclusive claim could involve a positive result in a control condition, e.g., "although the animal passed the test, its performance could be fully explained by a simple rule as shown in the control condition").

3. The text of the primary group-level statistical inference made supporting the main claim, if applicable

From the results section, or statistical analysis, coders copied the text of the group inference. Coders were asked to include the analysis that they thought was most central to the overall claim of the article, and to include any test statistics and *p*-values if they were present.

4. The text of the primary individual-level statistical inference made supporting the main claim, if applicable. If multiple individual-level inferences were made supporting the main claim, we coded the first one presented.

Similarly, coders were asked to copy the text of the individual inference, if present, from the results section or statistical analysis. Coders were asked to include the analysis that they thought was most central to the overall claim of the article, and to include any test statistics and *p*-values if they were present. If there were multiple individual-level inferences supporting a main claim, e.g., three animals passed a test, coders were asked to copy the first presented (that still supported the main claim).

5. The sample size

Coders were asked to report the number of animals recruited for the test.

6. The number of animals "passing" a test, if applicable

This was performed only for experiments that had an individual-level statement, coders recorded how many animals "passed" the test in question.

From the reported group-level and individual-level inferences, I then extracted the exact *p*-values from the texts of the group- and individual-level statistical inferences that were coded. For 44 experiments where non-exact values were reported, e.g., $p < .05$, we calculated them from the reported test statistics, if sufficient information was available. In four cases where an inequality was reported at a very low level, yet I did not have sufficient information to code the exact *p*-value, e.g. $p < .00001$ (N = 2) and $p < .001$ (N = 2), I included these as equalities in my analysis. We additionally planned to assess how researchers interpreted non-significant *p*-values; however, I decided that the number of non-significant *p*-values (16) was insufficient for any robust analysis.

5.2.3.Coding and Reliability

Coders were trained on four pilot studies, and any disagreement in this coding phase was used to refine the protocol. All 116 experiments of my final sample were double coded: I coded all of the experiments, and CT, DA, JF and JvdM acted as second coders. Where there was disagreement between the double-coded items, this was resolved by a third coder by reference to the same coding protocol (AV, LO and SP).[10] For all disagreements, the third coder's decision was used in the final dataset, and in the three cases where the third coder's decision disagreed with both the first and second coder, the third

[10] CT: Camille Troisi ; DA: Drew Altschul; JF : Julia Fischer; JvdM: Jolene van der Mescht; AV: Alizée Vernouillet; LO: Ljerka Ostojić; SP: Sarah Placì

coder's choice was retained after discussion with BGF. For the three variables that were coded by copying and pasting text, these were coded as agreeing if there was substantial overlap between the content of both coders, decided by BGF. This changed from the criteria I archived in our working methods, which was that the coders agreed if 50% of the text overlapped. I changed this criterion as the coders often varied in how much text they included, despite focusing on the same claim or inference.

5.3. Analyses

I first present descriptive data and visualisations of the following information: the distribution of species and groups from my sample, the sample sizes across the experiments, the frequency of positive and negative claims, the frequency of group-level and individual-level inferences, and *p*-value distributions for both group-level and individual-level inferences. I then use these data to explore the types of claims and strength of statistical inferences present in the animal physical cognition literature. I do the latter by qualitatively comparing the *p*-value distributions of the group-level and individual-level data with simulated distributions from researchers studying only true statistical effects with 80% power, 20% power, or studying zero true effects, i.e., 5% power. For these distributions, data were simulated from two normal distributions for each of the three sets of simulations. Population 1 had a mean of 50 and standard deviation of 5 for each, whereas Population 2 had a standard deviation of 5 for all different powers, but the means varied as follows: 80% power simulation, mean = 52.02; 20% power simulation, mean = 50.81; 5% power, mean = 50.

The difference between Population 1 and Population 2 was calculated to give the desired power for a two-tailed two sample *t*-test with n = 50 per group. 100,000 samples were then taken from each Population and compared to each other, and the *p*-values under .05 were plotted in Figures 11 and 12 alongside the *p*-values we sampled from the physical cognition literature. I made only qualitative conclusions from these data due to their non-independence and uncertainty about the theoretical *p*-value distributions under the different hypothetical scenarios (80% power, 20% power and 5% power). This uncertainty is commented on in the Discussion, as seeing the data first can help understand the limitations of the analysis.

5.4. Results

5.4.1.Coding Reliability

The first two coders agreed on 56 out of 63 (89%) claims from each paper's abstract, Cohen's K = .89, and agreed on 38 of 63 (60%) of their levels, i.e., whether the claims were positive, negative or inconclusive. This agreement rate was slightly lower than anticipated; however, all but two of the

disagreements occurred when one of the coders labelled a claim "inconclusive" and the other labelled it as either "positive" or "negative." Accounting for the ordinal structure of the data, in which a disagreement between "positive" and "negative" is more severe than a disagreement between "positive" or "negative" and "inconclusive", Cohen's $K_{weighted}$ = .47. Coders agreed on 45 of 75 (60%) group-level inferences, Cohen's K = .60, and 63 of 93 (68%) individual-level inferences, Cohen's K =.68. Sample sizes were coded equally in 103 of 116 experiments, Cohen's K = .89, and the number of animals passing each test was agreed in 56 of 94, Cohen's K = .59. All disagreements were then resolved by a third coder to produce the dataset for the analysis.

The inter-rater agreements for the group-level inferences, individual-level inferences, and claim levels were lower than anticipated. For the group-level and individual-level inferences, we performed my *p*-distribution analyses twice. My primary analysis was performed on the inferences that both coders agreed with, or that the third coder decided on in cases of disagreement. My robustness analysis used the inferences that both coders agreed on, and the inference that the third coder did not select in cases of disagreement (unless one of the original coders made a clear error, i.e., sometimes a coder would label an individual-level statistical test as a group inference, which BGF excluded from the robustness analyses). Perhaps more concerning was the lower inter-rater agreement for the claim level (60%). The following four reasons could explain this: i) my definitions being ambiguous and hence some claims being difficult to fit to them, ii) individual coders have response biases, iii) agreement on what the definitions mean but genuine disagreement on how to characterise the papers, and iv) typographical errors.

To investigate this further, I asked the original coders to recode all 63 claims (the claims in which both original coders agreed, or the claim which was decided by the third coder). Six coders (BGF, AV, CAT, JvdM, LO, SP) completed the re-coding. Between these six coders, the 15 inter-rater agreements, i.e., all possible pairings of coders, were, in percent: 51, 62, 64, 67, 68, 68, 68, 68, 68, 70, 71, 73, 76, 76 and 84. On average (median = 68%) this is slightly higher than the original 60%, and higher than chance (33%). This slightly higher agreement is likely due to the coders all having the same claim sentences to code, whereas in the original protocol, main claims were coded differently in 11% of cases. From this round of coding, I labelled the claims as positive, negative, and inconclusive if most of the coders (4 or more out of 6) chose one of the three categories. Claims in which at least 3 coders chose inconclusive and 3 or less chose positive or negative were labelled as inconclusive. Figure 8 visualises the inter-rater agreement; the 63 claims are presented sequentially along the x-axis, such that one column displays each coder's decision for that claim. The red dashed line in Figure 8 shows how these claims were separated (negative leftmost,

inconclusive centre, positive rightmost). My original coding (n = 2 coders and a third coder for disagreements) produced 14 negative, 10 inconclusive and 39 positive claims, and the second coding (n = 6 coders) produced 11 negative, 15 inconclusive and 37 positive claims. The overall agreement, concerning the final classification of claims of "positive", "negative" or "inconclusive", between the two rounds was 86%.

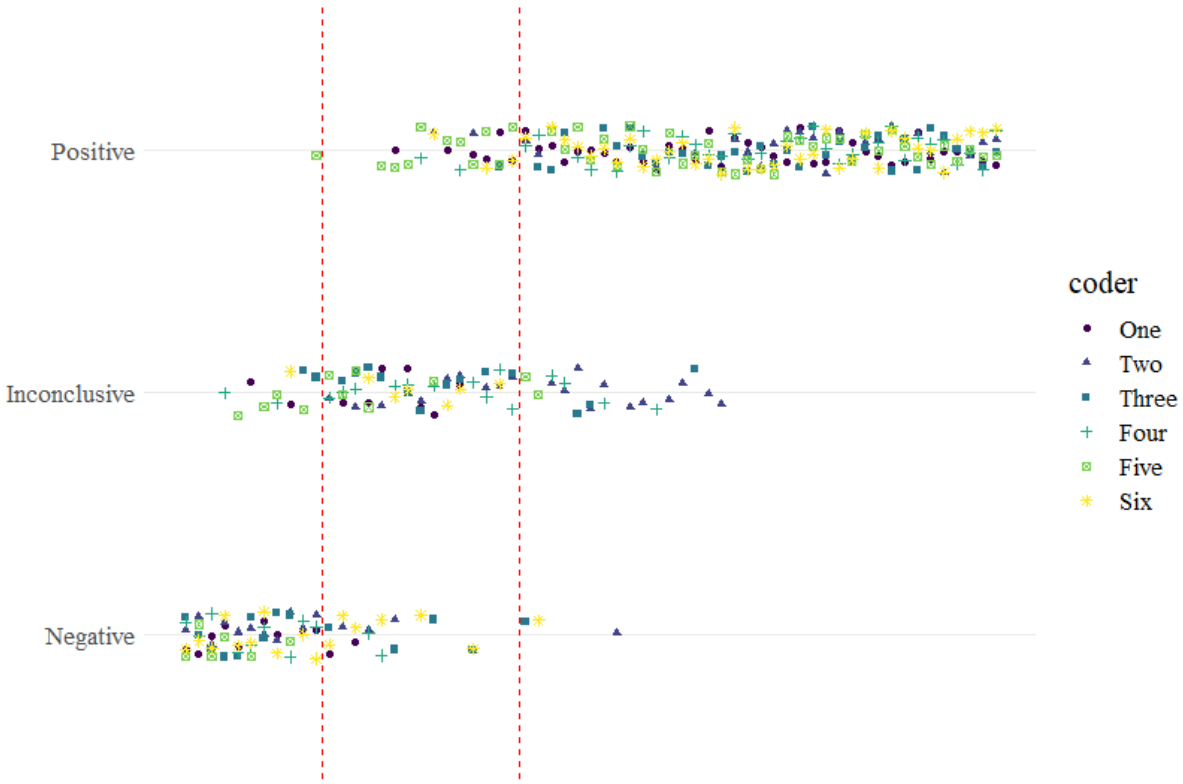


Figure 8: Inter-rater agreement when all 63 claims were re-coded by 6 coders using the original coding criteria. The 63 papers are presented sequentially along the x-axis and each coder's decision is presented on the y-axis (Positive, Inconclusive, Negative). Papers with most negative ratings are presented on the left-hand side, and papers with the most positive ratings are presented on the right. The red dashed lines indicate what category a particular paper was coded as when all the coders' responses were taken into account: papers were coded as 'Negative' (leftmost) or 'Positive' (rightmost) when at least 4 coders coded it as falling into that category, and papers were coded as 'Inconclusive' (centre) when the number of coders coding the paper as either positive/negative or inconclusive was equally split, (i.e., 3 coders in each category).

Figure 8 shows that most of the coders agreed on most of the claims, and when they did not, the disagreements centered on whether a claim should be labelled as inconclusive or positive, or inconclusive or negative. This suggests that a more continuous measure of claim levels might be useful in future research. Individual response biases also played a role: Coders One and Five (blue circles, green squares), preferred to code claims as more positive, whereas Coders Two and Four (blue triangles, blue crosses)

were more likely to respond with inconclusive rather than positive. While there was a good inter-rater agreement for many claims, there were some that split opinion across the six coders, which was likely due to ambiguity in the original definitions of “positive”, “negative” and “inconclusive” claims, as well as these not being natural categories for “claims”. To assess the effects of vague definitions on the findings, I provided the same 6 coders with more specific definitions and example sentences, and again asked them to recode the 63 claims. For this round, which occurred one to two weeks after the previous round, the instructions given to the coders were as follows:

“Positive and negative claims will make a general statement about the cognitive abilities, processes or behaviour displayed by the animals. Specifically:

Positive claims will include, but are not limited to, general statements about animals: having a certain cognitive ability; being able to pass a test; any claimed discovery of the processes animals use when passing a test (excluding “lower-order” processes); or confirmation of a general hypothesis. For example:

Our results suggest that…

> The animals understand physical causality
>
> The animals understood the causal structure of the task
>
> The animals are readily capable of passing the task
>
> Domestication has improved physical cognition in the animals
>
> The animals used X process to complete the task (where X process is not better categorized by the definitions for negative or inconclusive)

Negative claims will include general statements about the physical cognitive inability of animals, or their inability to pass a task and this not being caveated by an alternative explanation (such cases – where the inability is caveated – will typically be “inconclusive” and covered in a different category), or general evidence against a hypothesis. For example:

Our results suggest that…

> The animals do not understand physical causality
>
> The animals did not understand the causal structure of the task
>
> The animals might not be capable of passing the task

Suggesting that domestication has not improved physical cognition

Inconclusive claims will either not make strong epistemic statements but point to task specific confounds or task specific alternative explanations, OR they will report/refer to mixed evidence, i.e., some positive and some negative results OR a failure to confirm or disconfirm a hypothesis. For example:

Even though the animals passed the test, this likely did not require a causal understanding

Even though the animals passed the test, they likely achieved this through a simple rule

The animals may have understood some, but not all, of the physical properties of the task, and this was inconsistent between individuals

Using the specific definitions, inter-rater agreement was comparable to the original definitions used: 54, 56, 65, 65, 67, 67, 68, 70, 71, 71, 73, 73, 79, 86. However, the new definitions produced fewer positive (28) or negative (6) claims, and more inconclusive (29) claims (Figure 9). The overall agreement, concerning the final classification of claims of "positive," "negative" or "inconclusive," between this second round and the original coding 70%. This discrepancy is largely accounted for by the increase in inconclusive claims under the new definitions. Interestingly, coders' response biases were somewhat consistent across the two sets of definitions: Coders One and Five were again more likely to code claims as positive than the others, and Coder Two more likely to respond with inconclusive rather than positive.

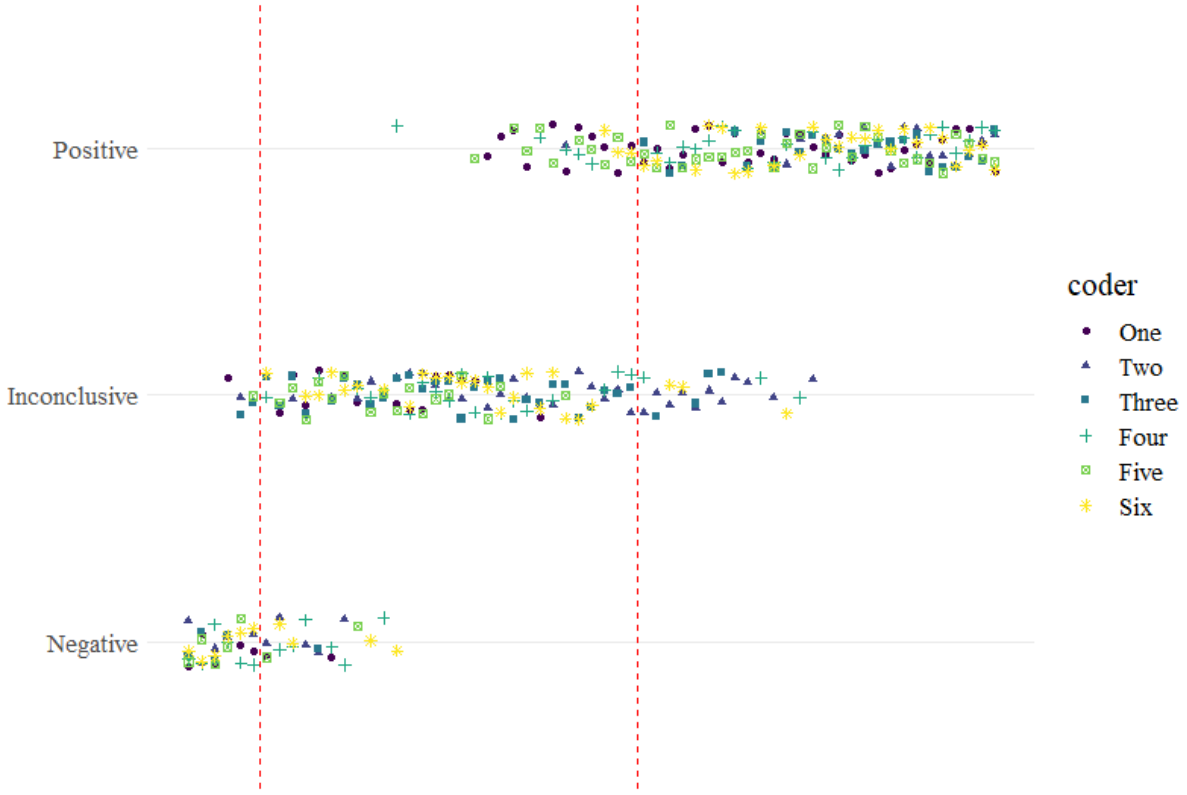


Figure 9: Inter-rater agreement when all 63 claims were re-coded by 6 coders using more specific coding definitions. The 63 papers are presented sequentially along the x-axis and each coder's decision is presented on the y-axis (Positive, Inconclusive, Negative).

Even though I was unable to increase the inter-rater agreement by providing more specific definitions, across the two rounds of coding, a relatively clear pattern emerged: some claims were clearly positive, some clearly negative and many intermediate. The intermediate claims, which I labelled "inconclusive" appeared continuously distributed. Some inconclusive claims were more "negative," and

others were more "positive," and coders tended to agree when this was the case. Examples of these claims are presented in Table *7*. Next, I present the results of my analyses.

Table 7: Examples of how claims were coded from our papers across two rounds of coding by 6 individuals.

Claim level category	**Example**
Positive	"The results indicate that kea are capable of assessing the spatial means–end relationships of this problem spontaneously and in a way that is comparable with primates." (Auersperg et al., 2009) "Our results showed that tool-use training enhances mean performance in the physical cognition domain, i.e. the understanding of spatial relations, numerosity and causality." (Tia et al., 2018)
Positive (Original Definitions); Inconclusive (Revised Definitions)	"Our subjects attended to at least two of the three tool features, although, as expected, the location of the hook was of paramount importance" (St Clair & Rutz, 2013)
Inconclusive	"These data are consistent with the idea that apes may possess some specific causal knowledge of traps but may lack the ability to establish analogical relations between functional equivalent tasks." (Martin-Ordas et al., 2008) "Both species [dogs and wolves] succeeded the visible displacement tasks but failed the invisible displacement problem" (Fiset & Plourde, 2013)

Negative (Original Definitions); Inconclusive (Revised Definitions)	"Bonobos did not demonstrate an understanding of contact but showed more individual variation, attending to the positions of the food, disk, and stick." (Helme et al., 2006)
Negative	"With the trap-tube task, we assessed whether the monkeys understood the cause-effect relation between their behavior and the outcome. The performances of the 4 subjects indicate that they did not take into account the effects of their actions on the reward." (Visalberghi & Limongelli, 1994)
	"Nevertheless, all 28 subjects failed to solve this task spontaneously, and showed no evidence of learning across 50 trials. Our results therefore call into question the earlier suggestion that dogs have, or can acquire, an understanding of the solidity principle." (Müller et al., 2014)

5.4.2.Sample sizes and biological units of interest

In our sample, the sample sizes ranged from 1 to 56, with a median of seven individuals (Figure 10). Eight of the experiments testing fewer than five individuals were transfer tasks. There was some evidence of species-specific differences in sample sizes, for example all corvid studies had a sample size of fewer than 10 individuals, whereas studies with canines and non-great ape primates sometimes had larger samples, for example over 50% of canine studies tested > 25 animals. Across the 116 experiments, 40 experiments included both group- and individual-level inferences in support of their claims, 28 used only group-level inferences and 48 only individual-level inferences.

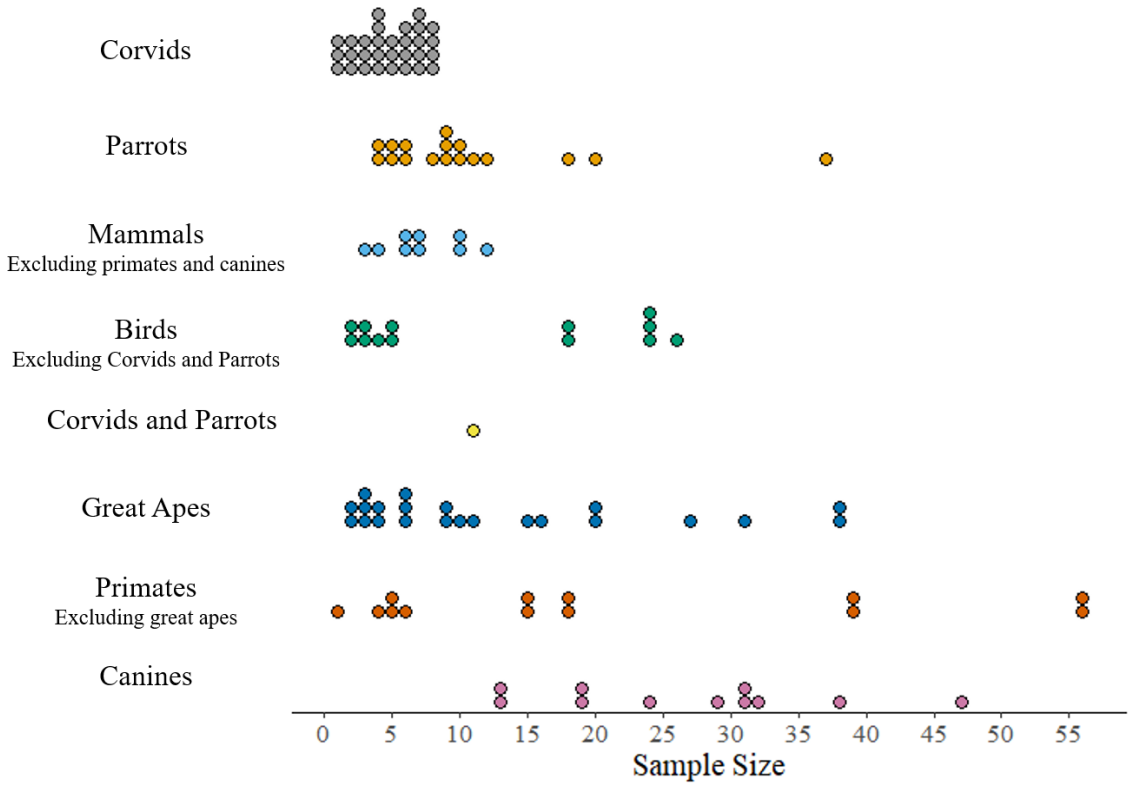


Figure 10: The distribution of the total sample sizes from each experiment of our sample of physical cognition experiments. NB: data points are not independent as some of the sample sizes are from experiments from the same papers, and some of the papers are from the same laboratories that tested the same sets of captive animals. The group "Parrots and Corvids" comes from a single paper that studied both groups.

5.4.3.Claims

From our two coding rounds in which six individuals coded each claim, I found that between 28 and 37 of the 63 papers (44 - 59%) made positive claims. Between 6 and 11 (10 - 17%) papers made negative claims, and between 15 and 29 (24 - 46%) papers made inconclusive claims (for the definitions used for each claim, see the methods section, and for examples see Table 7). Table 8 shows how these figures vary across groups, and although our sample size is small, corvids and great apes have the largest proportion of positive claims.

Table 8: The number of each type of claim (positive, negative or inconclusive) made in the sample of physical cognition papers, by study group. The first value comes from the first round of coding using the original definitions (n = 6 coders), and the second value from the second round of coding using the

tighter definitions (n = 6 coders). The group “Parrots and Corvids” comes from a single paper that studied both groups.

Group	Positive	Negative	Inconclusive
Great Apes	7; 4	1; 0	4; 8
Primates excluding great apes	4; 4	1; 1	2; 2
Canines	4; 2	1; 1	2; 4
Mammals excl. primates and canines	3; 2	1; 0	1; 3
Corvids	12; 9	0; 0	3; 6
Parrots	4; 4	4; 2	1; 3
Parrots and corvids	0; 0	0; 0	1; 1
Birds excluding corvids & parrots	3; 3	3; 2	1; 2

5.4.4.P-values

From the 68 group-level inferences, I had sufficient information to extract 58 exact *p*-values, of which 46 were < .05. From the 88 individual-level inferences, I had sufficient information to extract 49 exact *p*-values, of which 41 were < .05. The density distributions of these significant *p*-values are plotted in Figure 11, alongside simulated distributions from research with 80% power to detect a true effect (“80% power simulation”, uppermost plot), research with 20% power to detect a true effect (“20% power simulation”, second from top), research where all null hypotheses were true (“False positive simulation”, third from top).

Both the group- and individual-level statistical inferences have the largest density of *p*-values between 0 and .01, providing evidence of correct rejections of H0. This pattern is clear for the individual-level inferences, which is consistent with research performed at relatively high power. Similar distributions, albeit with slightly larger p-values for the individual inferences, were observed when I re-

performed the robustness analysis using the data from the second coder, when they disagreed with the primary and third coders, and are plotted in Figure 12.

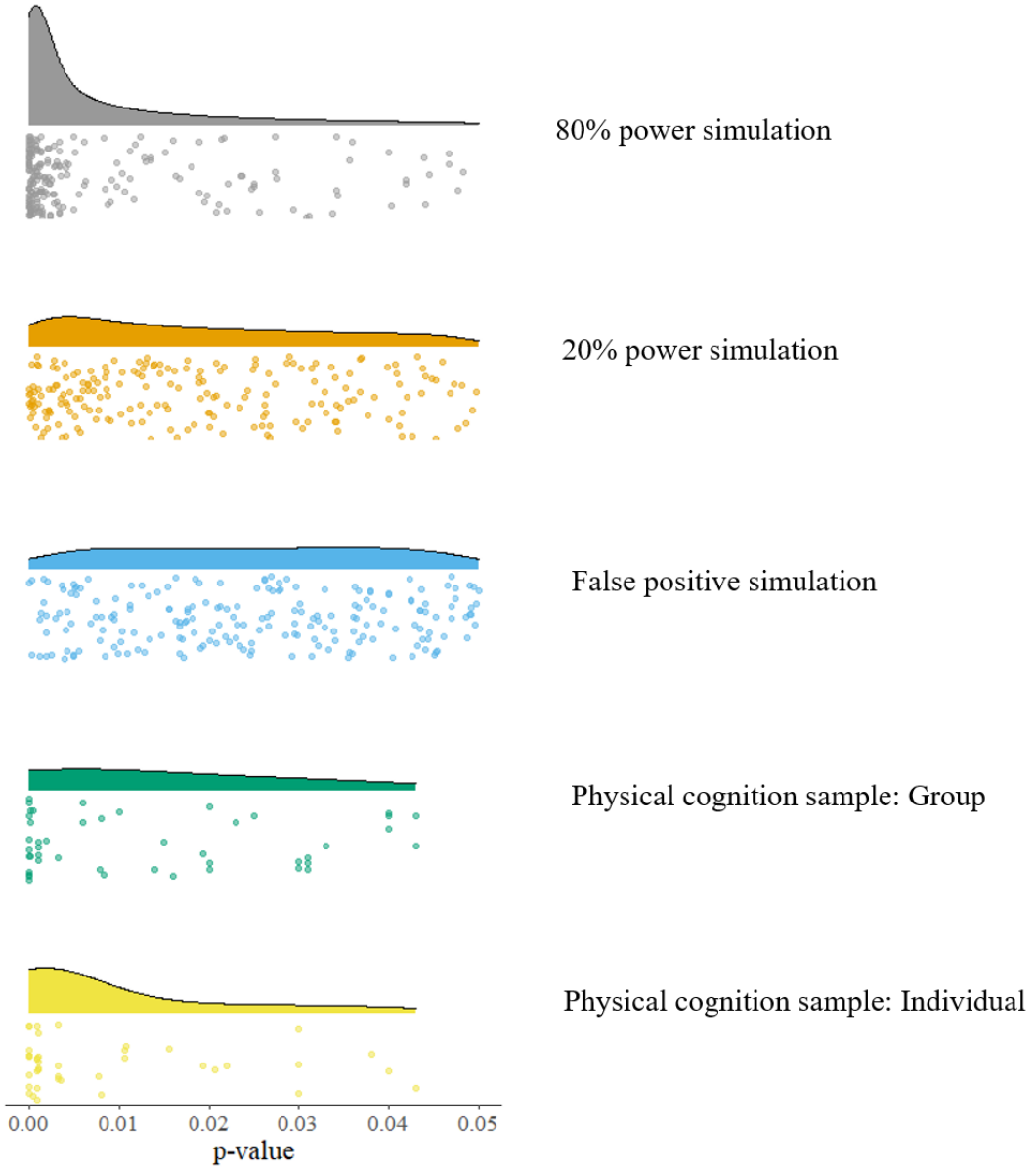


Figure 11: Raincloud plot of significant p-values resulting from simulated research with H1 always correct and research performed with 80% power (uppermost plot), 20% power (second panel from top), and H0 always correct (third from top), the group-level inferences from our sample of physical cognition research (N=46, fourth panel from top) and the individual-level inferences from our sample of physical cognition research (N=41, bottom panel). The simulations and their density distributions contain 100,000 p-values each, however for clarity a random sample of 200 raindrops are presented underneath the density plot.

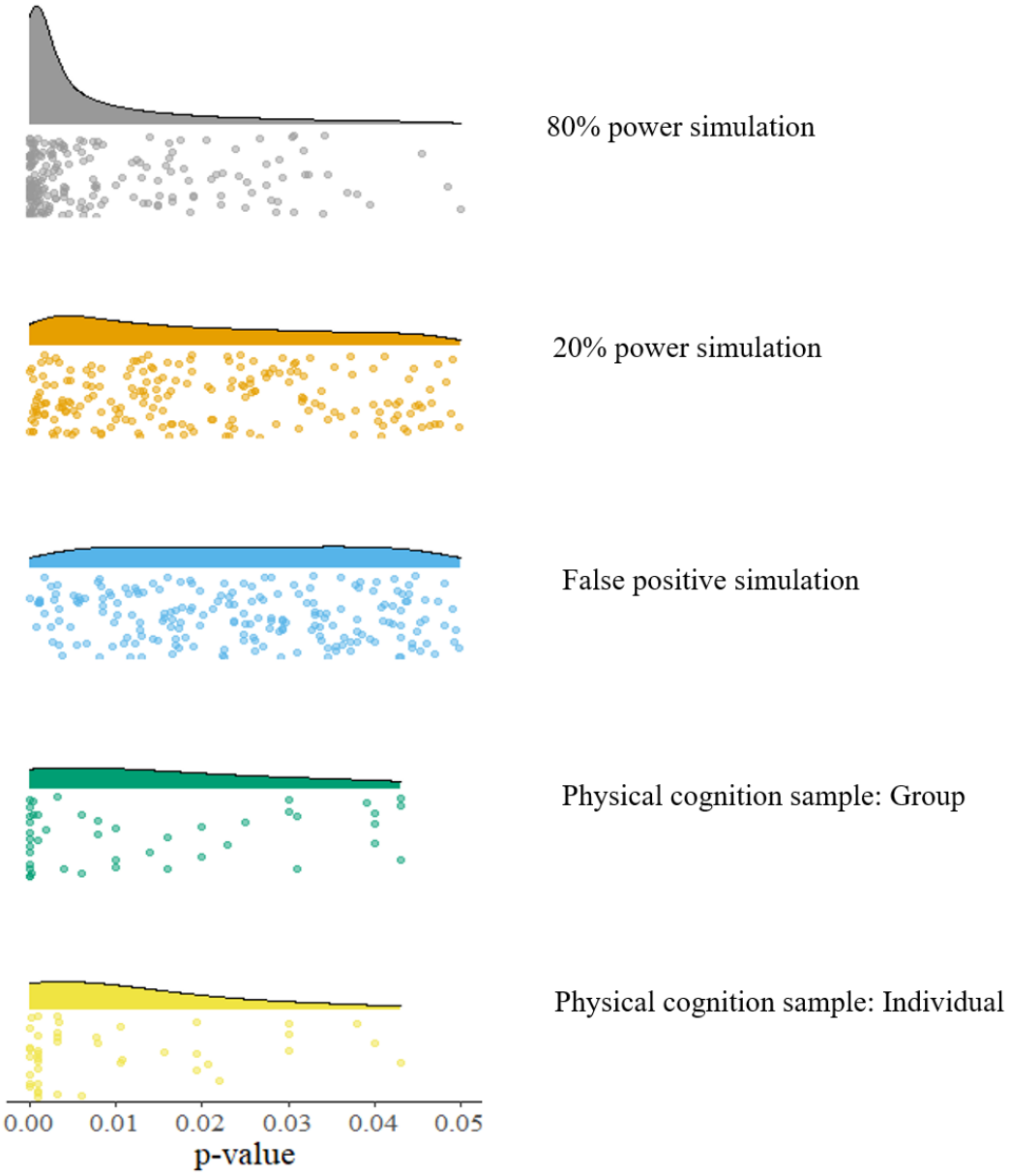


Figure 12: Raincloud plot of significant p-values resulting from simulated research with H1 always correct and research performed with 80% power (uppermost plot), 20% power (second panel from top), and H0 always correct (third from top), the alternative group-level inferences from our sample of physical cognition research (N=45, fourth panel from top) and the alternative individual-level inferences from our sample of physical cognition research (N=43, bottom panel). The simulations and their density distributions contain 100,000 p-values each, however for clarity a random sample of 200 raindrops are presented underneath the density plot.

I performed two further exploratory analyses with the group-level *p*-values. The upper panel of Figure 13 displays the p-values as a function of sample size, with a red dashed line at $p = .05$, and the area $.01 < p < .05$ shaded. This graph provides weak evidence for larger sample sizes to produce smaller *p*-values, however some very small *p*-values were still reported from studies with small sample sizes. The lower panel of Figure 13 is a histogram of the frequency of different p-values across the range 0 – .50. The number of *p*-values drops appreciably just above .05, suggesting that either some results just above this threshold are unpublished or that the *p*-value has been deflated to below the threshold.

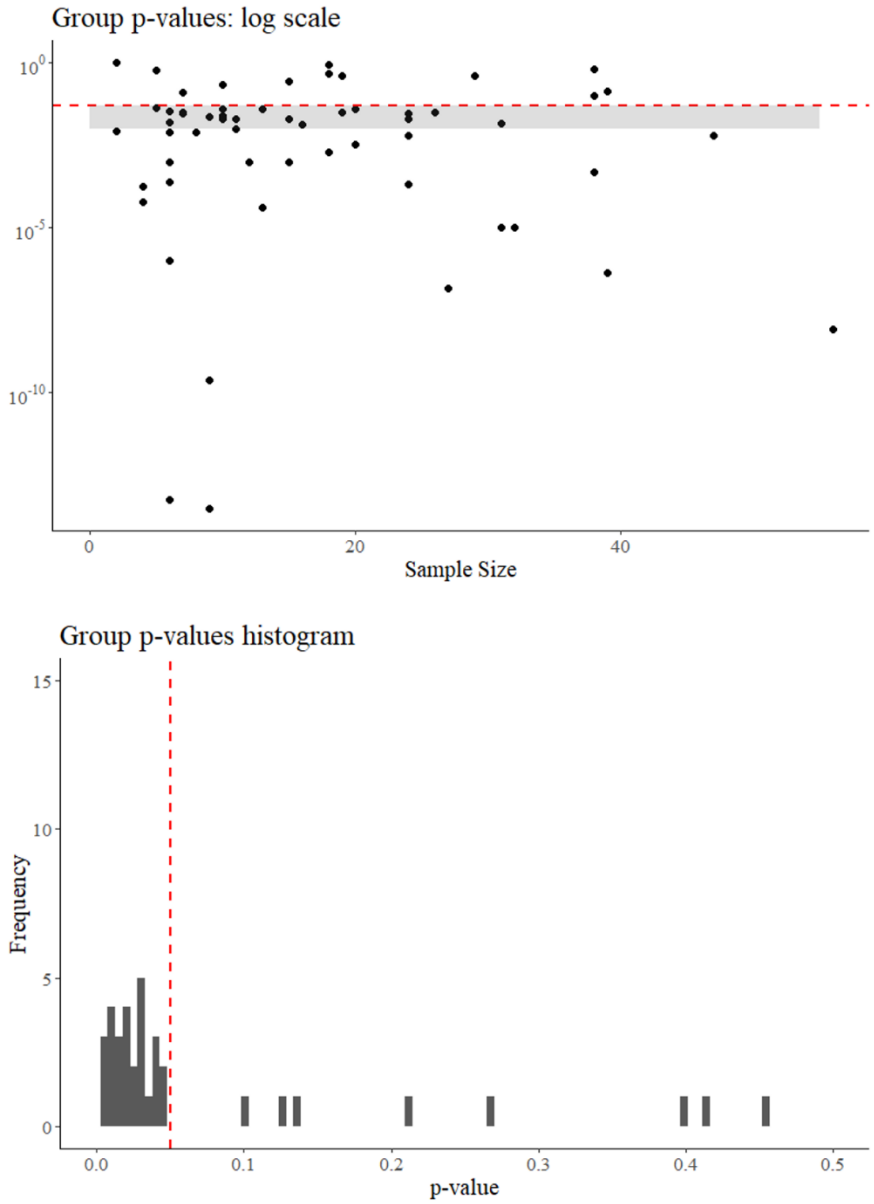
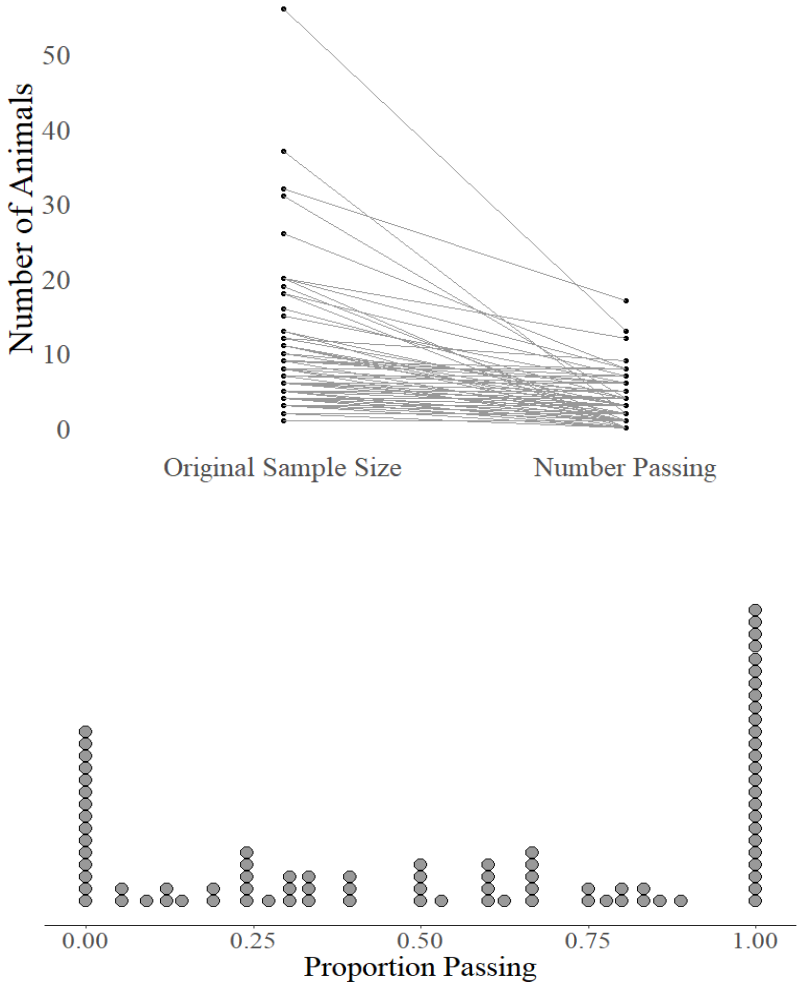


Figure 13: The distribution of group p-values from our sample as a function of sample size (upper panel), and frequency (lower panel). On the upper panel, the x-axis gives the sample size and the y-axis the size of the

reported p-value. The red dashed line denotes a p-value of .05 and the grey shaded denotes the area in which p-values fall between .01 and .05. The lower panel shows a histogram of the frequency of different p-values across the range of 0 – .50 (only 2 p-values fell between .50 and 1 so I excluded this range for clarity). The red dashed line denotes a p-value of .05.

5.4.5.Animals "Passing" Each Test

Of the 88 experiments making individual-level inferences, we were able to code the number of individuals who "passed" the experimental test from 87 experiments. In 15 experiments, zero individuals passed the test. In the remaining 72 experiments in which individuals passed the test, the maximum number of individuals passing a test was 17 and the median was 3. The number of individuals passing each test, and the corresponding original sample size are plotted in Figure 14 (top panel). The proportion of animals passing the experimental tests varied substantially between experiments, from 0 individuals in 15 experiments to 1 (all individuals) in 25 experiments. The median proportion of animals passing an experimental test was 0.6 (bottom panel).



5.5. Discussion

In this study, I collected data on the types of claims and statistical inferences used in a sample of physical cognition papers. Our sample contained some, but far from all, task formats used in animal physical cognition research: 14 variants of the trap tube task, 13 other tool use tasks, 11 means-end, contact and sting pulling tasks, 9 object choice, object tracking and object permanence tasks, 5 access tasks, 3 studies used the physical tests from the Primate Cognition Test Battery, 3 support problems, 2 examined how birds choose nest building material, 1 paper tested dogs' understanding of solidity, 1 investigated tool manufacturing in crows and 1 investigated both means-end understanding and tool use in lemurs. Moreover, these tasks were not randomly sampled. As I used the search terms "physical cognition," "folk physics" and "trap-tube" variations, specific findings or authors may be overrepresented – for example, some authors may be more likely to use the phrase "folk physics" than others, and some not included at all (although no author was represented more than five times across the sample). Because of these sampling biases, the exact numbers for many of the measures are unlikely to be accurate representations of the field as a whole; however, the data provide patterns that are relevant to some general features of the literature.

5.5.1. Statistical Design and Inference

5.5.1.1. Sample Sizes and Biological Units of Interest

In line with previous reports of sample sizes in animal cognition research and comparative

Figure 1: The number of animals passing each experimental test, and the corresponding original sample size from our sample of physical cognition research (top panel). The proportion of animals passing the test in each experiment from our sample of physical cognition research (bottom panel).

sted
eral
015)

tested 47 dogs, and Joly et al. (2017) tested a total of 39 macaques of four different species. The sample also did not include some experiments with very large sample sizes, such as Herrmann et al. (2007) who tested 107 chimpanzees and 32 orangutans on the Primate Cognition Test Battery, because the abstract, title or keywords of the paper did not include the terms "physical cognition," "folk physics" or "trap-tube."

Across the 116 experiments, 40 experiments included both group- and individual-level inferences in support of their claims, 28 used only group-level inferences and 48 only individual-level inferences. Hence, most experiments (88) focused, at least in part, on individual animals' performances, consistent with suggestions that this is the appropriate level of analysis for many psychological theories (Barlow & Nock, 2009; Skinner, 1956; Smith & Little, 2018). However, focusing on the individual animal does not

always resolve problems with small sample sizes. If only a small proportion of individuals will "pass" a test of physical cognition, then studies with small sample sizes risk missing such individuals altogether. I found that, in 13 experiments, between 1% and 25% of the sample passed the test at hand (Figure 14), showing that small sample sizes can still be a concern, even if researchers focus their statistical analyses on the individual.

5.5.1.2. Publication Bias and Statistical Reliability

Depending on the definition of positive, inconclusive, and negative claims, between 44% and 59% of the sample made positive claims, between 24% and 46% made inconclusive claims, and between 10% and 17% made negative claims. There was therefore no universal bias towards clearly positive claims in our sample and, given that our sample represented a host of major animal cognition journals, this finding might generalize to animal cognition research. However, I also found a noticeable drop in *p*-values just above the significance threshold for group-level inferences, which is, by definition, a marker of publication bias. Taken together, this suggests that, while it has been possible to publish negative and inconclusive results in physical cognition research, publication bias may still be an issue for the field. Importantly, the presence of publication bias likely interacts with researchers and research groups, with some being more likely to publish negative findings than others. Any literature-wide analysis, like the present study, does not account for this interaction, and therefore may miss patterns of publication bias. However, a large proportion of positive and inconclusive claims, such as with corvids and great apes in our sample, does not mean that publication bias is present; it is also consistent with the animals genuinely performing well on most tasks.

Rather than providing strong evidence about publication bias and statistical reliability in physical cognition research, our descriptive analysis illustrates one type of method that could provide such data in the future. In our sample, the *p*-value distribution from the individual-level inferences was consistent with relatively high-powered tests at the individual-level. This was less of the case for the group-level inferences, which appeared more consistent with a mixture of high and low-powered tests and suggests that some effect sizes may be overestimated, and hence difficult to replicate (Farrar et al., 2020; Hedges, 1984). I could not perform formal *p*-curve analyses with our data because *p*-curve analyses, to be effective, require i) independent data points; whereas I included multiple experiments from the same studies, and ii) the same hypothesis to be tested; whereas our sample included many different tests in many different species.

5.5.2. Methodological Concerns and Future Research into Statistical Reliability and Publication Bias in Comparative Cognition

My analysis indicates that even though negative and inconclusive reports have been published in animal physical cognition research, publication bias likely still influences the validity of the published literature. However, more importantly, the challenges we faced in the current study highlight general problems that studies attempting to quantify publication biases and statistical reliability in animal cognition will face.

5.5.2.1. Sampling and Level of Analysis

The generalisability of my findings is limited by our sampling method - a keyword search for terms sometimes used by researchers studying physical cognition ("physical cognition" and "folk physics"), which I biased further through adding the search term "trap tube." Future research may wish to generate more comprehensive search terms to cover the population of studies in a research area (Chapter 6) or take a random sample of articles from across animal behaviour and animal cognition journals (Chapter 7).

5.5.2.2. Definitions and Reliability

The inter-rater reliabilities of our measures were low, specifically when researchers assessed the level of a claim, which I attempted to categorize as positive, inconclusive, or negative, and the text of the main statistical inferences supporting these claims. Currently, it is unclear to us what "good" inter-rater reliability scores will be for our questions, and while it should be possible to increase the inter-rater reliability through greater training and feedback for coders, or through more prescriptive definitions, I believe if such measures were used to generate near 100% inter-rater reliabilities, this would reduce the measures' validities (Penders et al., 2019).

5.6. Summary and Next Steps

In this chapter, I analysed 116 animal physical cognition experiments from 63 published journal articles. Sample sizes were small on average (median = 7), but some studies with larger sample sizes were also observed. Depending on the definitions I used to categorize positive, inconclusive, and negative claims, between 44% and 59% of our sample made positive claims, between 24% and 46% made inconclusive claims, and between 10% and 17% made negative claims. These data suggest that there was no universal bias towards positive results in our sample. Nevertheless, there was a drop in the frequency of *p*-values reported above the significance level, suggesting the publication bias is still an issue in the field. Our *p*-distribution analysis suggests that researchers are often finding true statistical effects at the individual-level; however, some published group-level effects in animal physical cognition research may

be overestimated. In the following chapter, I extended on the methods developed in this chapter to perform a systematic review and risk-of-bias assessment across nearly the entire corvid social cognition literature.

6. Chapter 6: A systematic review to assess risk-of-bias in studies of corvid social cognition

Corvids are suggested to be “one of the most intelligent groups of animals on the planet” (Taylor, 2014, p. 361). A swathe of data have been published in support of this claim, from field observations (Goodwin, 1956), to neuroanatomy and neural activation data (Balakhonov & Rose, 2017; Emery & Clayton, 2004; Kirschhock et al., 2021; Lefebvre et al., 2004; Nieder et al., 2020; Olkowicz et al., 2016), arguments based on evolutionary theory (Clayton et al., 2007; Emery & Clayton, 2004; Emery et al., 2007), and, most prominently, behavioural experiments (detailed in the following paragraph). Collectively, these data are often used to claim that corvids might be capable of sophisticated cognition, including second-order reasoning, and that these abilities rival those of the great apes (Clayton, 2012; Emery & Clayton, 2004; Taylor, 2014).

Corvids are claimed to be particularly intelligent in social cognition (Clayton et al., 2007; Emery et al., 2007). Corvids have an enlarged nidopallium caudolaterale (Olkowicz et al., 2016), analogous to the mammalian pre-frontal cortex, and social structures that are often complex and competitive, and require high-quality relationships to be maintained. Because of this, corvids are oft cited as having both the neural scaffolding and historical evolutionary pressures that would lead to complex social cognition being selected for (Emery, 2006; Emery & Clayton, 2004, 2008). This claim has been explored intensively with captive corvid species across several labs worldwide. For example, corvid cache protection and pilfering strategies might require flexible cognition to maximize access to food resources, both when protecting their own caches from others and pilfering the caches of others (Bugnyar & Heinrich, 2006; Dally, Clayton, et al., 2006). And, in co-operative scenarios, pair-bonded male Eurasian jays might be able to adjust their food sharing behaviour in-line with their partner’s desires and choices (Crosby et al., 2020; Ostojić et al., 2013). Many of these findings have been motivated by the concept of theory of mind. For example, corvid social behaviour has been suggested to involve visual perspective taking (Bugnyar et al., 2016), experience projection (Emery & Clayton, 2001), knowledge attribution (Bugnyar, 2011; Dally, Emery, et al., 2006), and desire attribution (Ostojić et al., 2013). Elsewhere, corvids have been claimed to perform transitive inference to predict the dominance of others (Paz-y-Miño, Bond, Kamil, & Balda, 2004), spontaneously co-operate on tasks (Massen et al., 2015), display emotional contagion (Adriaense et al., 2019), and display elements of mirror-self recognition (Clary & Kelly, 2016).

However, given the concerns raised so far in this thesis about how scientific research and publication practices can promote overestimated and overgeneralised findings (Gelman & Carlin, 2014; Higginson & Munafò, 2016; Ioannidis, 2005; Nissen et al., 2016; Smaldino & McElreath, 2016; Yarkoni, 2019), and given recent null findings and replication failures of studies of corvid social cognition (Amodio, Brea, et al., 2021; Amodio, Farrar, et al., 2021; Brecht et al., 2018; Crosby, 2019), a systematic and critical review of corvid social cognition research is due. However, rather than summarising the results of previous studies, this systematic review aims to search and map the landscape of the corvid social cognition literature and then apply the methods developed in Chapter 5 to build a preliminary risk-of-bias assessment for the field. Thus, the review has the following objectives:

1) To describe the demographics of corvid social cognition research: Where are studies performed, on which species and on which topics?
2) To outline the sample sizes used in corvid social cognition research and how or whether they were justified
3) To collect the claims papers on corvid social cognition make and categorise these as each other positive, negative or inconclusive (*sensu* Chapter 5), as an indirect measure of publication bias
4) To collect and investigate the *p*-values distributions of the focal hypothesis tests for each paper, as an indirect measure of evidence strength and publication bias
5) To extract the frequency of direct replication, experimental blinding and inter-observer reliability calculation across the field
6) To extract the frequency of accessible data and code across the field

6.1. Methods

6.1.1.Eligibility criteria

Articles were included in the review if they fit all of the following criteria:

1. Article Type: Original research article, published in an academic journal or book
2. Language: English language
3. Species: Paper has at least one study on corvids and this study is on social cognition
4. Intervention: Researchers performed a manipulation or intervention in the environment, i.e., not passive observations

5. Topic: Social cognition, defined as the study of how corvids acquire, store or process social information, or information that researchers have provided to copy some element of a social scenario. The primary focus of at least one part of the paper should be cognitive rather than behavioural, although it is recognised that there is no clear distinction

6.1.2.Information sources

Scopus was searched between the 7th and 15th April 2021. No other databases were searched, meaning that two non-Scopus-indexed journals that could have published articles on corvid social cognition, *Animal Behavior and Cognition* and *International Journal of Comparative Psychology*, were not represented in this search.

6.1.3.Search Strategies

To define social cognition, LO, solicited key words associated with the study of social cognition from all collaborators on this project and several further experts in the field. Using these key words to search Scopus, I identified a total of 7,835 papers (see Table 1 of the Appendix for the exact terms and search).

6.1.4.Selection Process

6.1.4.1. First screen

While on Scopus, I performed the first screen of papers by deselecting papers that were clearly off topic for export based on their titles and abstracts. These included studies from different disciplines or not on corvids. A total of 909 candidate papers were exported for the second screen.

6.1.4.2. Second screen

The second screen comprised of a calibration phase (25% of articles) and a double-blind phase (75%) and based on the titles, abstracts and – if required - full texts of the articles. Two coders (BGF and LO) rated whether each article contained at least one study fulfilling our inclusion criteria. In the calibration phase, LO was not blind to my answers, and in the double-blind stage LO was. I then reviewed any disagreements and made a decision as to whether to include or exclude the article, or – in a case of uncertainty - entered it into the third screen.

6.1.4.3. Third screen

Articles over which there was uncertainty in the second screen were entered into the third screen. Here, 7 coders voted whether each paper fit the inclusion criteria or not. The majority vote was taken as the final decision in each case.

6.1.4.4. Extraction and quality control

During data extraction and quality control, coders had the opportunity to flag articles or studies that were included as not fitting the inclusion criteria. These decisions were reviewed by myself or ME after quality control.

6.1.5.Data Extraction

Twelve coders with expertise in animal cognition, and often corvid social cognition research, participated in data extraction (myself, AB, AV, CT, CW, EGP, EL, JA, KB, LO, ME, SP)[11]. I added a topic area to each paper, and coders were asked to indicate topics with which they were most familiar. Coders were then non-randomly assigned 1/12 of the papers according to their topics of expertise. Each coder either had participated in the physical cognition project of Chapter 5, or was trained on several papers of the same project. During this training, and during an additional pilot phase, a coding manual was developed for use during the extraction process. This manual, available at https://osf.io/3aznq/, detailed how the data items would be extracted for every social cognition study in each of the selected articles. These data items are detailed in Table *9*.

Table 9: Data items extracted in the review

Category	Data Item	Data Extracted	Extracted *per*
Demographics	Laboratory	The surname of the principal investigator of the laboratory most closely associated with the laboratory	Article
	Country	The country in which the research was conducted	Article

[11] AB: Amalia Bastos; AV: Alizée Vernouillet; CT: Camille Troisi; CW; Claudia Wascher; EGP: Elias Garcia-Pelegrin: EL; Ed Legg; JA: Jessie Adriaense; KB: Katharina Brecht; LO: Ljerka Ostojić; ME: Mahmoud Elsherif; SP: Sarah Placì

Sample	Species	The common and Latin names of the species tested	Study
	Sample Size	The number of animals recruited for the study	Study
	Sample Size Justification	Whether the total sample size was justified, and if so, what the justification was	Study
Claims	Title Claim	The title of the article *if* the title could be construed as a claim	Article
	Title Claim Level	Whether the title claim was positive, negative or inconclusive[1]	Article
	Abstract Claim	The main theoretical claim of the article in the abstract	Article
	Abstract Claim Level	Whether the abstract claim was positive, negative or inconclusive[1]	Article
Result	Result Text	The text of the result that most closely matched the main claim in the abstract	Study
	p-value	The p-value from the result text most closely related to the article's main claim	Study
Design	Replication	Whether the study was a direct replication, and whether this was in the same species or a different species to the original study	Study
	Direct Interaction	Whether the experimenter was directly interacting with the animals during the study, where direct interaction was defined as any possibility of the experimenter producing Clever-Hans-like effects	Study

	Experimental Blinding	Whether the experimenter was blinded to the conditions when performing the study	Study
	Measurement	How the behaviour was measured	Study
	Inter Observer Reliability (IOR)	Whether any inter-observer reliability was calculated	Study
	IOR Blinding	Whether the second coder was blinded to the hypotheses, conditions or results of the first coder when coding	Study
Data and Code Accessibility	Open Data	Whether data were fully (all raw data necessary to reproduce the analyses), partially (some raw data available, but not enough to reproduce the analyses) or not available	Study
	Open Code	Whether any code that would be able to reproduce the analysis was shared	Study
Citations	Web of Science Citations	Scopus automatically extracted the number of citations *per* Web of Science	Article

[1]The definitions for positive, negative or inconclusive claims were adapted from Chapter 5: Positive: "a claim that suggests the presence of an ability in the animal(s), a (novel) effect, the animal(s) "passed" a test or a claim of a discovery about the similarities or differences between two groups of animals' cognition."; Negative: "a claim that suggests the absence of an ability in the animal(s), the absence of novel effect, or that the animal(s) "failed" a test, and that this failure was due to the animals not having the required cognitive or behavioural abilities."; Inconclusive: "a claim that does not strongly suggest that the animal(s) have or do not have a certain ability or that an effect does or does not exist. Instead, it will point to inconsistencies in its results, confounding factors or any other explanations for the animals' performances, which means the data is not diagnostic for the ability in question. This also includes mixed evidence about a claim."

6.1.6.Double blind extraction

I double-blind coded 38 articles (25%). These studies were randomly selected from those coded by the other coders using the function 'sample' in R, such that 4 articles from 10 other coders were double-blind extracted. Two of the 40 studies were identified as not fitting the inclusion criteria and thus excluded.

6.1.7.Quality control

128 (84%) of articles further went through a non-blind quality control procedure. Here, a second coder reviewed the data extracted from each article. If the quality controller identified a mistake, they classified this as a major disagreement, and if the quality controller disagreed but was uncertain, for example in the case of borderline claims, they classified this as a minor disagreement. Quality controllers were not required to go back to the main text to verify each statement, rather they checked whether, for example, the extracted result made sense given the extracted claim. However, for seven of the data items (Abstract Claim, Direct Interaction, Experimental Blinding, IOR, IOR blinding, Open Data and Open Code), quality controllers did return to the full text to ensure no relevant information was missed. Myself and ME then reviewed all disagreements highlighted during quality control (50% of articles each) and made a final decision on what entered the final dataset, returning to the original article if necessary. Finally, LO – one of the most experienced team members - performed the quality control checks for all laboratory and country decisions in the dataset.

6.1.8.Analysis

All data are presented descriptively and visualised where appropriate.

6.2. Results

6.2.1.Study inclusion

The Scopus search produced 7835 articles, of which 909 were retained following the first title/abstract screen. Next, 478 duplicates were removed, and of the remaining 431 articles, 266 were excluded during the second screen. During the initial, non-blind stage of the second screen (25% of articles), agreement between myself and LO was 94.4%. For the remaining 75% of articles, myself and LO agreed on 88.8% of inclusion decisions, giving an overall agreement rate of 90.4%. Twelve articles were entered into the third screen, in which seven of the coders voted on whether or not an article should be included for further extraction. Of these, a further eight articles were excluded, and four more articles were excluded during

data extraction due to not being an interventional study. These data are presented in Figure 15. The total number of articles on social cognition in corvids where the authors made any intervention was 153, some of which contained multiple studies, such that there was a total of 226 different studies.

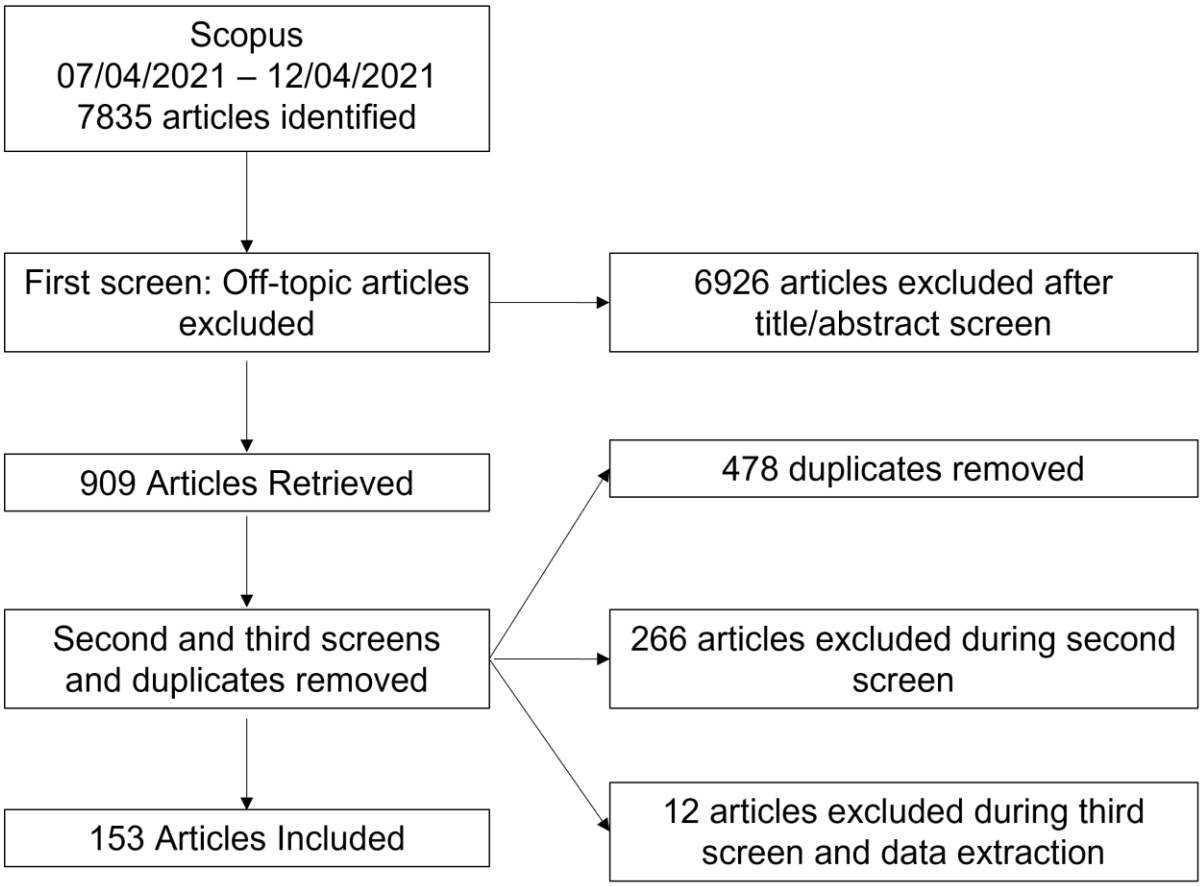


Figure 15: A Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRIMSA) diagram of the screening process

6.2.2. Double-blind data extraction and quality control

The inter-rater agreement for the double-blind coding and quality control are detailed in Table 10 and Table 11, respectively. For most measures, inter-rater agreement was good, with the following exceptions. In the double-blind coding, coders only extracted 56% of the same sentences for the result text corresponding to the main claim of the article (and because of this only agreed on 62% of p-values as these were usually within the result text). This is likely due to several potential results mapping onto the main claim of the paper, of which the extractors had to choose one. However, in quality control, the quality controllers agreed with 95% of the extracted results, suggesting that in most cases original coders were choosing justifiable results given the article's main claim. Inter-rater agreement was also low for the direct interaction and inter-observer reliability blinding items, both in the double-blind and the quality

control stages, which reflects a surprising level of ambiguity in how interaction and blinding are reported in papers.

Table 10: Inter-rater agreement following double blind extraction

Data Item	N agree	N disagree	% agree
Laboratory	43	7	86
Country	46	4	92
Species	47	3	94
Sample Size	36	14	72
Justification of Sample Size	47	3	94
Title Claim	29	9	76
Title Claim Level	14	2	88
Abstract Claim	29	9	76
Abstract Claim Level	29	9	76
Result Text	28	22	**56**
p-value	31	19	62
Replication	42	8	84
Direct Interaction	33	17	66
Experimental Blinding (if agreed on Direct Interaction)	7	0	100
Measurement	46	4	92
Inter Observer Reliability (IOR)	44	6	88
IOR Blinding	12	5	71
Open Data	43	7	86
Open Code	49	1	98

Table 11: Number of items verified or commented on during quality control

Data Item	N verified (%)	Minor comments	Major comments

Species	197 (99%)	1	0
Sample Size	176 (89%)	15	7
Sample Size Justification	188 (95%)	7	4
Title Claim	118 (92%)	6	2
Title Claim Level	117 (91%)	9	5
Abstract Claim	110 (86%)	13	4
Abstract Claim Level	117 (91%)	7	76
Result Text	188 (95%)	8	2
p-value	184 (93%)	9	5
Replication	182 (92%)	13	2
Direct Interaction	162 (82%)	22	14
Experimental Blinding	173 (87%)	13	12
Measurement	174 (88%)	14	10
Inter Observer Reliability (IOR)	190 (96%)	4	4
IOR Blinding	189 (95%)	5	4
Open Data	188 (95%)	9	1
Open Code	193 (97%)	1	4

6.2.3.Demographics

6.2.3.1.1. Laboratory

Forty-seven different laboratory groups were identified across sixteen different countries Figure 16. However, during quality control, LO and I discussed doubts about the utility of the laboratory data, as multi-lab collaborations and shared authorships made these data difficult to interpret. They are therefore not presented systematically, although the data are used occasionally in the species and topics discussion.

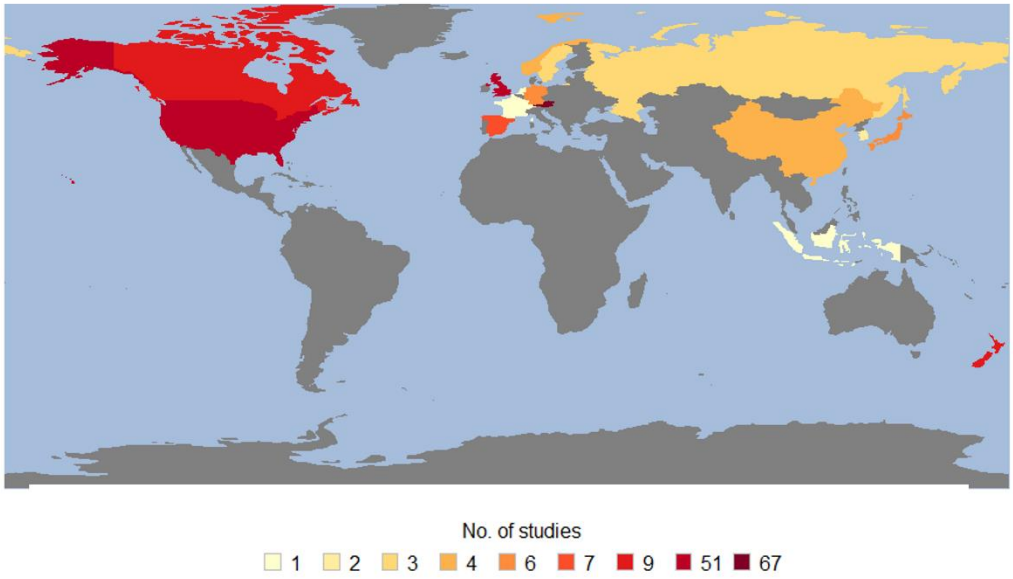


Figure 16: The distribution of interventional studies of corvid social cognition

6.2.3.1.2. Species and topics

A total of 22 different species were tested across the 153 articles, 16% of all 136 recognised corvid species (Table 12). Ravens were the most often represented species, tested in over 1/3rd of articles, and California scrub-jays and Western jackdaws were also tested in over 1/6th of all articles. For many topics and species combinations, a single research group was either the only research group studying them or had published over three-quarters of all research. Because of the uncertainty around the laboratory data, this is only a tentative conclusion, but the example of the laboratory I performed this research in is given for illustrative purposes: The Comparative Cognition Lab at the University of Cambridge published 18 of 20 papers on cache protection, perspective taking or theory of mind in California scrub-jays, and all 9 of such studies in Eurasian jays. The topics detailed in Table 12 are non-definitive and used for basic description only.

Table 12: The species represented in our dataset of interventional studies on corvid social cognition. N studies refers to the number times a species was represented in the sample, and Topics details which topics were studies in them, with the number in brackets being the number of studies within that species and topic.

Species	N studies	Topics
California scrub jay *Aphelocoma californica*	25	Social response to death (5) Cache-protection, perspective taking or theory of mind (20)
Florida scrub jay *Aphelocoma coerulescens*	4	Social learning (2) Cache-protection, perspective taking or theory of mind (2)
Transvolcanic jay *Aphelocoma ultramarina*	1	Individual recognition (1)
White-throated magpie-jay *Calocitta formosa*	1	Social learning (1)
American crow *Corvus brachyrhynchos*	8	Social learning (5) Individual recognition (1) Cache-protection, perspective taking or theory of mind (2)
Common raven *Corvus corax*	54	Affiliation (2) Cooperation or prosociality (12) Communication (5) Contagion (3) Individual recognition (1) Relationship recognition (1) Social learning (6) Cache-protection, perspective taking or theory of mind (23)
Hooded crow *Corvus cornix*	3	Mirror response (2) Transitive inference (1)
Carrion crow *Corvus corone*	12	Communication (1) Individual recognition (5) Transitive inference (1) Cooperation or prosociality (4) Mirror response (1)

Rook *Corvus frugilegus*	10	Cooperation or prosociality (3) Contagion (3) Individual recognition (2) Cache-protection, perspective taking or theory of mind (2)
Large-billed crow *Corvus macrorhynchos*	6	Individual recognition (4) Mirror response (1) Social learning (1)
Western jackdaw *Corvus monedula*	25	Cooperation or prosociality (3) Communication (1) Individual recognition (3) Mirror response (2) Relationship recognition (1) Social learning (7) Cache-protection, perspective taking or theory of mind (3) Transitive inference (5)
New Caledonian crow *Corvus moneduloides*	11	Cooperation or prosociality (5) Mirror response (2) Social learning (4)
House crow *Corvus splendens*	1	Mirror response (1)
Blue jay *Cyanocitta cristata*	5	Cooperation or prosociality (2) Individual recognition (2) Social cue use (1)
Azure-winged magpie *Cyanopica cyanus*	6	Mirror response (4) Cooperation or prosociality (2)
Eurasian jay *Garrulus glandarius*	11	Social learning (2) Cache-protection, perspective taking or theory of mind (9)
Pinyon jay *Gymnorhinus cyanocephalus*	10	Cooperation or prosociality (3) Social bonding (2) Cache-protection, perspective taking or theory of mind (3) Transitive inference (2)

Clark's nutcracker *Nucifraga columbiana*	10	Cooperation or prosociality (1) Mirror response (4) Cache-protection, perspective taking or theory of mind (5)
Siberian jay *Perisoreus infaustus*	1	Social learning (1)
Eurasian magpie *Pica pica*	4	Individual recognition (2) Mirror response (2)
Multi-species *Aphelocoma wollweberi* *Aphelocoma californica* *Cyanopica cyanus* *Cyanopica cyana* *Corvus monedula* *Corvus frugilegus* *Corvus corone* *Corvus corax* *Corvus macrorhynchos* *Corvus moneduloides* *Corvus corone* *Corvus corax* *Gymnorhinus cyanocephalus* *Nucifraga columbiana* *Perisoreus infaustus*	18	Cooperation or prosociality (4) Mirror response (2) Social interaction and attention (2) Social learning (8) Transitive inference (2)

6.2.4.Sample size

6.2.4.1.1. Sample size

The median sample size across all experiments was 9 (min: 2, mean: 13.75, max: 167).

6.2.4.1.2. Sample size justification

The sample size decision was coded as justified in only 14 (6%) of experiments. Of these, eight were justified by the constraints researchers faced, e.g., testing all available animals, and for the other six, statistical power was mentioned.

6.2.5.Claims

6.2.5.1.1. Title claim level

Ninety-two of the 153 articles (60%) had titles that coders interpreted as claims. Of these, 74 (80%) were coded as positive, 7 as inconclusive (8%), and 11 as negative (12%).

6.2.5.1.2. Abstract claim level and citations

Of the 153 articles, the abstracts main claim was coded as positive in 118 (77%), inconclusive in 19 (12%), and negative in 24 (16%) cases. Figure plots the number of citations each article received by year of publication and by claim type. Year of publication was a strong predictor of citation count ($F(1, 158) = 58.8$, $p < .0001$), whereas there was no statistically significant relationship between claim type and citation number after controlling for year ($F(2, 158) = 1.81$, $p = 0.17$). Nevertheless, in each year, the most cited paper had a positive claim, and all papers with over 50 citations had positive claims. It seems likely that positive claims are cited more frequently than negative claims, but I lacked the sensitivity in our sample to detect this.

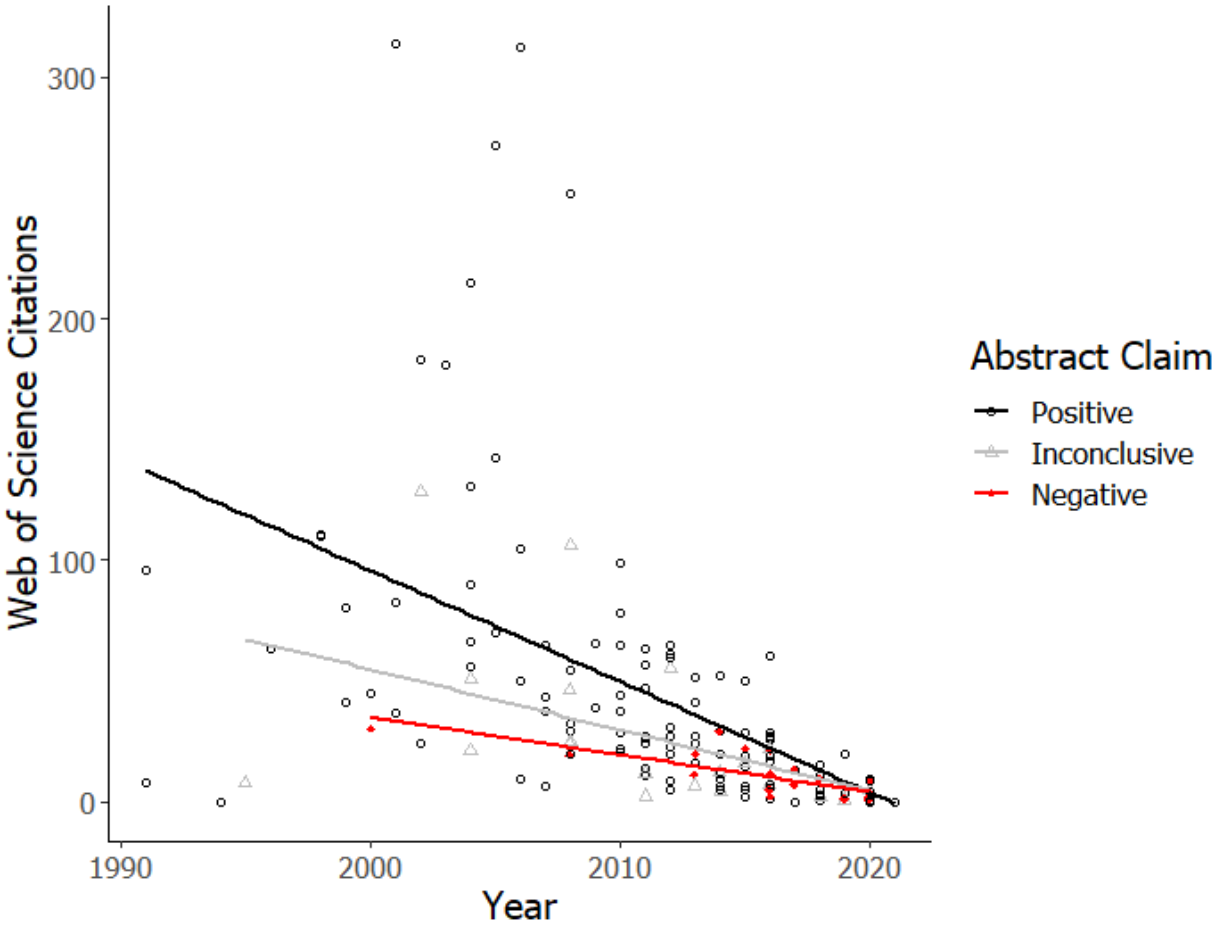


Figure 17: Citation number of articles by year and claim type

6.2.6.Evidence

6.2.6.1. P-values

One-hundred and thirty-four of the 223 experiments reported results with exact *p*-values. The distribution of these are plotted in Figure 18. There is a clear decrease in the number of *p*-values just above 0.05 compared to those that are just below, with a ratio of 5.25:1 of *p*-values from 0.03-0.05:0.05-0.07 (see Chapter 7 for a longer discussion on interpreting this drop in *p*-value density). A further 63 experiments reported *p*-values as inequalities, and these are plotted in Figure 19.

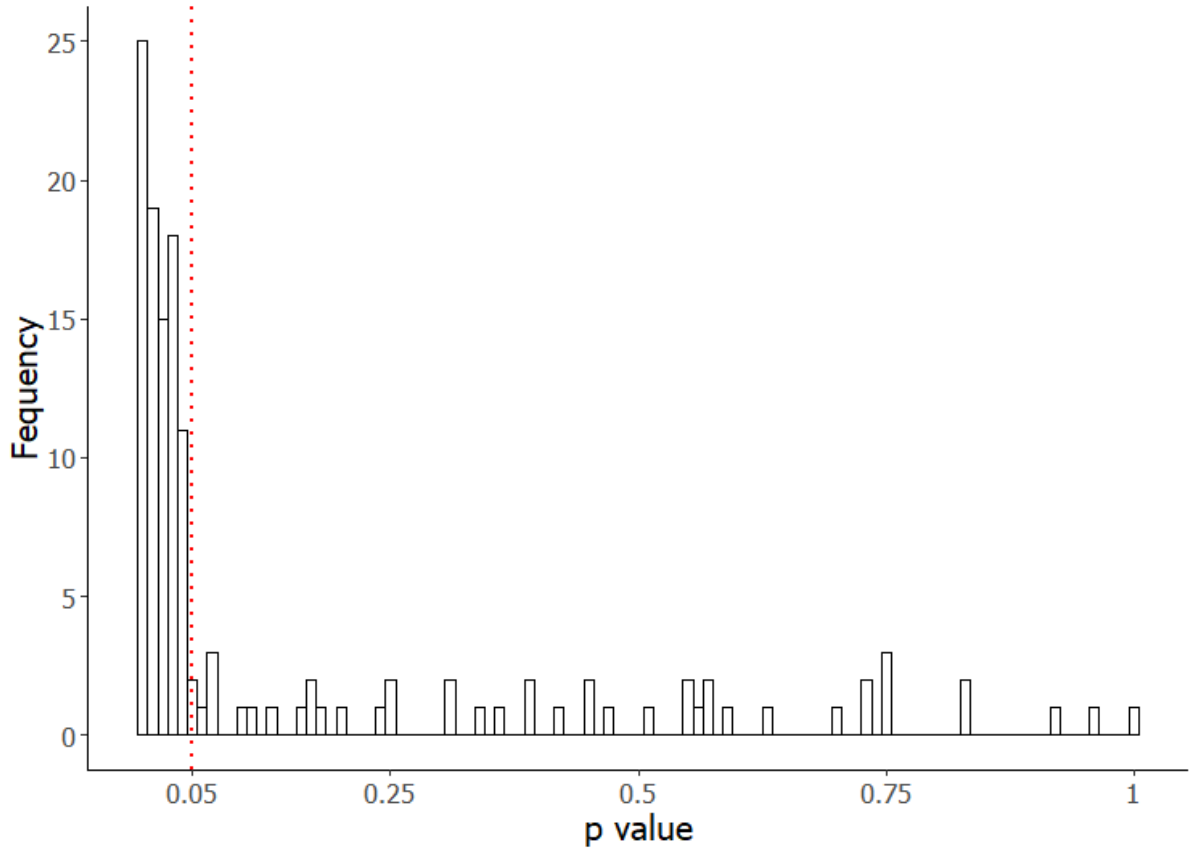


Figure 18: p-value distribution of the exact *p*-value most related to an article's main claim across 134 studies of corvid social cognition.

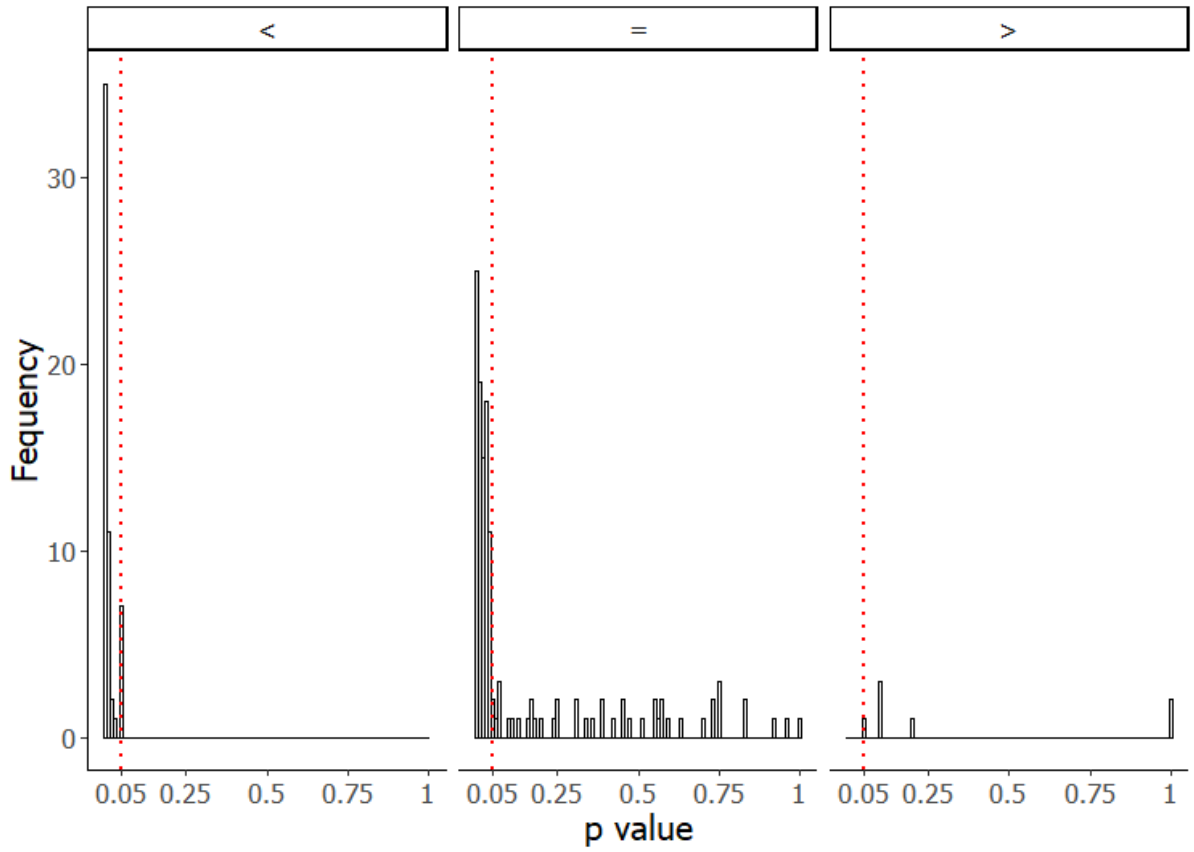


Figure 19: p-value distribution of the p-values most related to an article's main claim across 223 studies of corvid social cognition: Left panel, p-values reported as the inequality <; central panel, exact p-values; right panel, p-values reported as the inequality >.

6.2.6.2. Sub-group analyses

I had originally planned to perform sub-group analyses of the p-value distributions by topic area. However, upon exploring the data for the theory of mind/cache protection topic (Figure 20) – the largest subgroup - it was evident that the number of studies per subgroup was too small to facilitate a meaningful analysis.

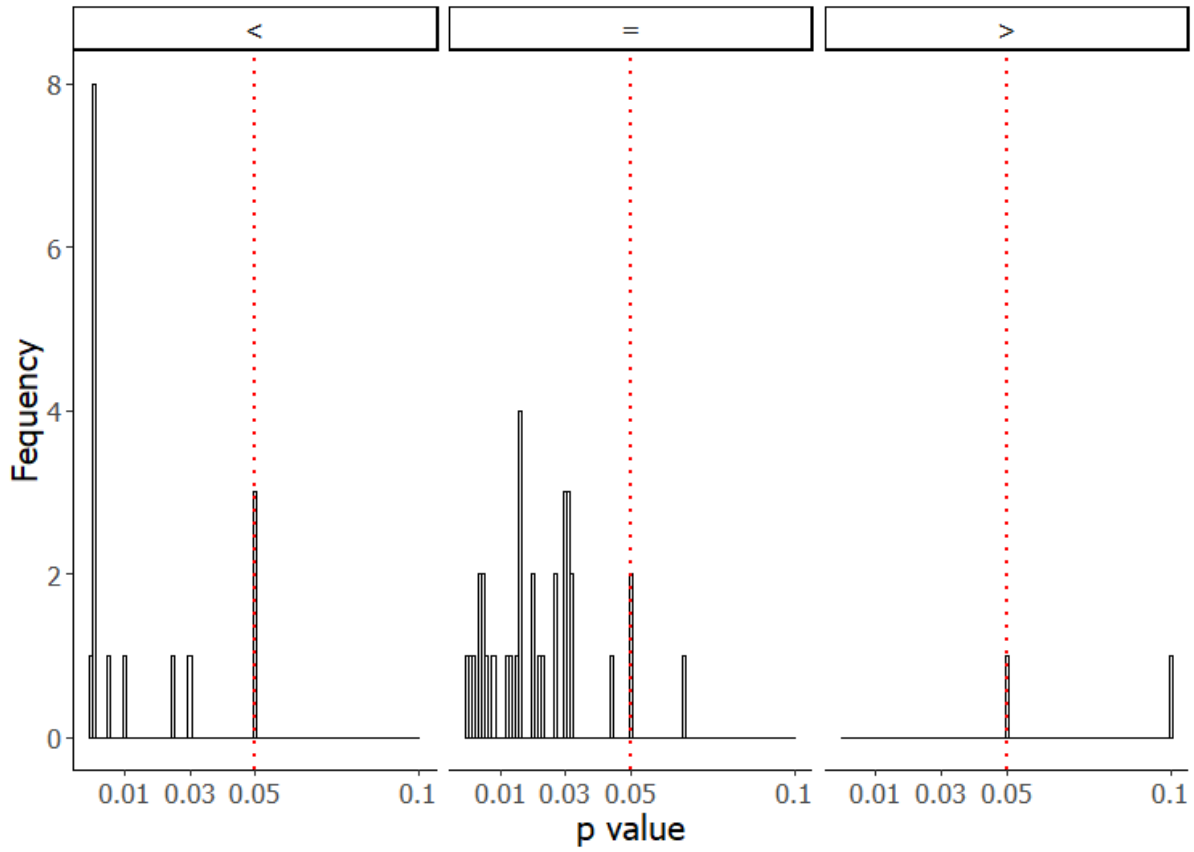


Figure 20: p-value distribution of the p-values most related to an article's main claim across 52 studies of corvid social cognition. For clarity, 12 p-values larger than 0.1 were not plotted. Left panel, p-values reported as an inequality '<'; central panel, exact p-values; right panel, p-values reported as an equality '>'.

6.2.7.Design

6.2.7.1. Replication

Of the 226 studies, 56 were identified as replications (25%). Thirty-six of these (16% of the total sample), were replications with a different species, in which the same paradigm has been used in a different species, and 20 (9% of total sample), were within-species replications, i.e., true "direct" replications. Figure 21 plots the rate of replication across years.

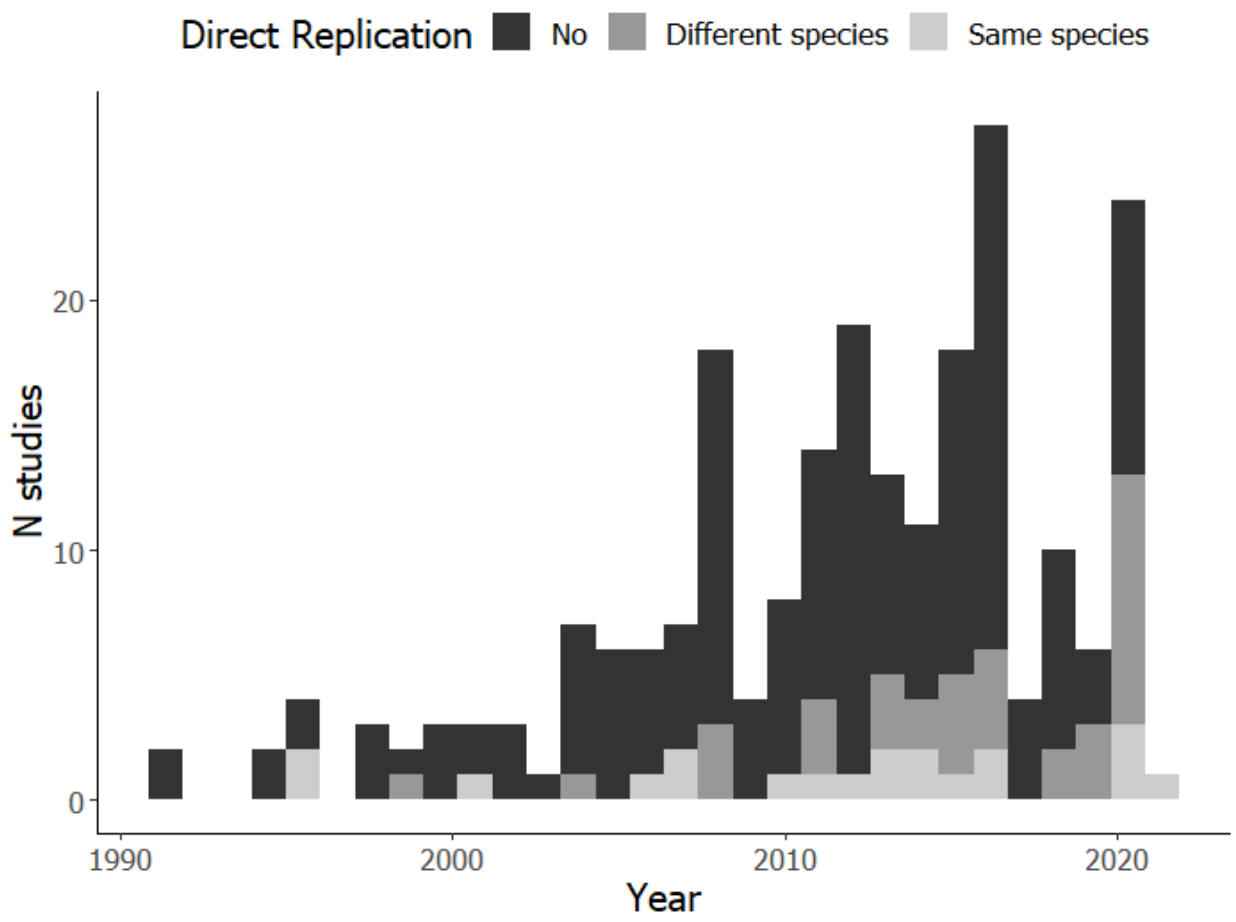


Figure 21: The rate of direct replication studies in corvid social cognition research

6.2.7.2. Experimental Blinding

Experimenters directly interacted with animals in a manner that could produce experimenter effects, if blinding was not used, in at least 81 studies (36%). Experimenters did not directly interact with animals during data collection in 98 studies (43%). Surprisingly, for the remaining 47 (21%) studies, methods were insufficiently described such that it was unclear to three experts (the original coder, quality controller and final reviewer) whether or not the experimenter was in direct interaction with the animals or not, in a manner that could bias the results.

When experimenters were in direct interaction with the animals, they were reported as blinded in only 9 of the 81 studies (1%). When it was coded as unclear whether the experimenters were in direct interaction, they were reported as blinded in only 7 of the 47 studies (15%).

6.2.7.3. Inter-observer reliability

The following methods were used to score behaviour: coding off video recordings (104 studies, 46%); direct observations (28 studies, 12%), indirect observations (17 studies, 8%); automatic observation (e.g., touchscreen, 14 studies, 6%); and mixed approaches (35 studies, 15%). Again, it was unclear to three experts what measurement technique was used in 27 (12%) of the coded studies.

A second observer was used, and inter-observer reliability (IOR) calculated, for 83 studies (37%). These second observers were reported as blinded in any fashion for 35 studies (40% of all studies that calculated IOR). More specifically, 14 studies (17%) reported that their second observer was blinded to the hypotheses of the study, 10 (12%) that their second observer was blinded to the conditions they were viewing, and only 2 (2%) that the second observer was blinded to the results of the first observer.

6.2.8.Data and code accessibility

6.2.8.1. Open data

Full data were openly accessible for 29 studies (13%), and some raw data – but not enough to reproduce the analyses – were available for a further 18 (8%) of studies. Data were inaccessible for the remaining 75% of studies. Figure 22 plots the data availability of the coded studies by publication year.

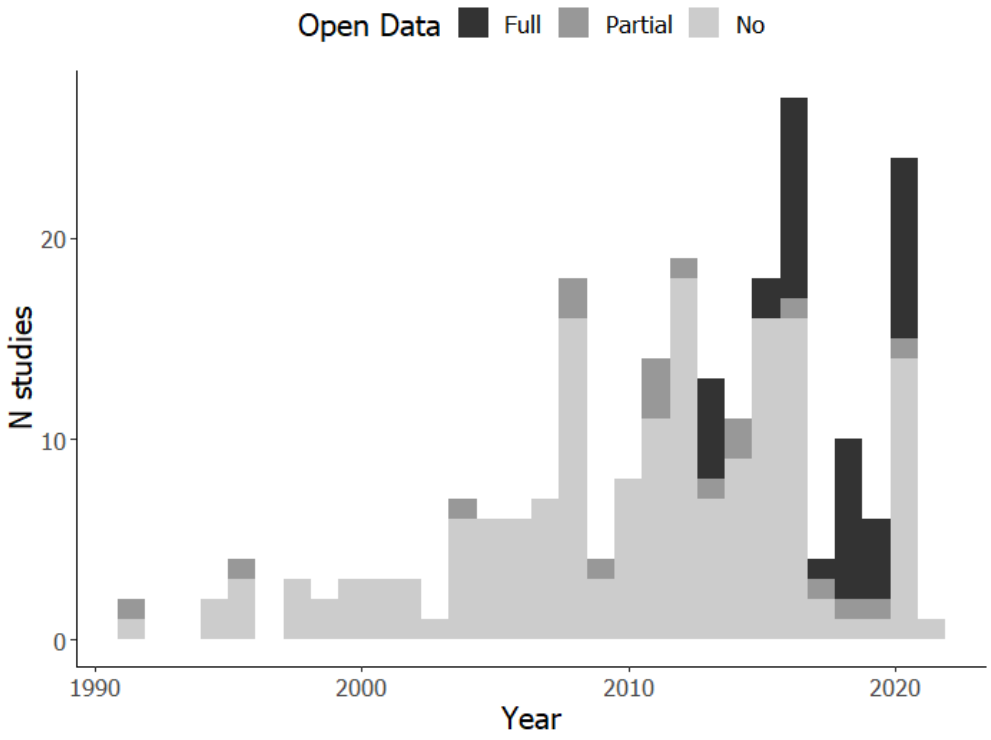


Figure 22: Data availability of corvid social cognition studies across publication year.

6.2.8.2. Open code

Analysis code was openly available for 15 (7%) of studies.

6.3. Discussion

This study provides a preliminary map and risk-of-bias assessment of interventional studies of corvid social cognition. Studies across the field may be considered at high risk-of-bias because of, i) the siloed nature of the field, ii) small sample sizes (but see Smith and Little (2018)), iii) positive claims dominating the literature but a *p*-value distribution indicating many negative results have not been published, and iv) the lack of reported blinding procedures for both experimenters and second observers. Moreover, the computational reproducibility and robustness of analysis strategies in the field can rarely be assessed due to historically low rates of data and code sharing. I now discuss each of these in turn.

While the analysis was not formalised, our study provides some evidence to support the notion that the research field is siloed – often with a single group publishing the majority of papers on a single species-topic combination, although there are exceptions. Sample sizes in corvid social cognition research, with a median of 9, are consistent with those reported elsewhere in animal cognition research (Chapter 5; Many Primates et al., 2019). Notably, sample sizes were rarely justified, with only 2.7% of papers discussing statistical power when detailing their sample. This is concerning as it suggests researchers have not been formally considering the risk of false negative results when designing studies, something that might be especially common with small sample research if many trials are not used (Chapter 2).

This study provides good evidence of a publication bias against negative results in corvid social cognition research. According to our coding scheme, the literature primarily consists of positive claims (80% of title claims and 77% of abstract claims), which was coupled with a clear drop in *p*-values above the 0.05 threshold (Figure 19). It seems most likely that there are three components to the corvid social cognition literature, i) clear true positive results with a high *a priori* likelihood (associated with the smallest *p*-values), ii) uncertain results at high risk of effect size overestimation (associated with *p*-values just under 0.05), and iii) a considerable number of unpublished negative results.

Within-species direct replication studies (9% of all studies) occurred at a rate slightly higher than those found for human psychology (1.6% pre-2012 Makel et al., 2012) but near identical to primate cognition research between 2014 and 2019 (8.7%, Many Primates et al., 2019). While this does not constitute habitual replication (nor is habitual direct replication an efficient long-term strategy, especially if publication bias can be minimized (Coles et al., 2018; Field et al., 2019; Halina, 2021; Isager et al., 2021)),

at least some replication studies have been published in the field. Nevertheless, the recent replication failures of (Amodio et al., 2021; Crosby, 2019) highlight the continued short-term need for direct replication studies of previously published findings in corvid social cognition research.

Concerningly, the rate of articles reporting that their experimenters during data collection (12.5%) and second coders (16%) during observation were blinded was low, and difficult for us to extract. Blinding has consistently been recognised as key to minimising bias in animal experiments (Beran, 2012; Holman et al., 2015; Tuyttens et al., 2014, 2016), but rarely reported: Burghardt et al., (2012) reported that less than 10% of animal behaviour articles from five journals reported either experimenter or inter-rater blinding, and Kardish et al., (2015) reported a blinding rate of 13.3% in ecology, evolution and behaviour. While our figures were low and are extremely concerning, they might slightly underestimate the actual rate of blinding due to errors in reporting. In line with this, when second coders were reported as blinded, it was rarely made explicit what they were blinded to – whether this was the hypotheses, conditions, or results of the first coder. Similarly, it is possible that some authors may not report blinding as they assume that readers would infer appropriate and common blinding procedures have been followed – however, such a lack of reporting would make it difficult to reproduce or critically assess the studies in question. Reporting guidelines, such as the Animal Research: Reporting of In Vivo Experiments (ARRIVE) guidelines (Sert et al., 2020), could facilitate stronger reporting of blinding procedures in animal cognition research.

Of similar concern was the low rate of data and code availability in the field, which sets a theoretical maximum for computational reproducibility without contacting the original authors. That data and code sharing was low across all years is unsurprising given data sharing mandates are relatively new and this finding is in-line with previous research from other fields (e.g., Culina et al., 2020; Minocher et al., 2020). Nevertheless, the conclusion remains that most analyses in corvid social cognition cannot be readily reproduced, and the data are unavailable for large meta-synthesis projects. Retrospectively, researchers could seek to archive past data, and prospectively make datasets and code openly available as far as ethically possible.

While this study highlights general concerns with the current state of the corvid social cognition literature, it does not mean that any individual study has been identified as biased. Studies must be assessed on an individual basis, something that is especially important given the level of heterogeneity in the field. For example, it makes no sense to decrease our confidence in a report of transitive inference in Siberian jays, just because an independent study of theory-of-mind in carrion crows did not use a blinding procedure (hypothetical examples). However, overall, this study does highlight the continued need for

caution when interpreting findings in corvid social cognition research. Biasing factors, such as publication bias and experimenter effects have not been clearly avoided, and truly independent research seems rare for most species-topic combinations (see Chapter 4). Finally, this study highlights four key areas in which corvid social cognition research can improve in the future: i) replication of uncertain findings, ii) publishing of non-significant results, iii) implementing of blinding procedures, and iv) improved reporting of blinding procedures.

The current Chapter (on corvid social cognition) and Chapter 5 (on physical cognition) extracted a range of data to describe and investigate specific subfields of animal cognition research. Next, in Chapter 7, I will provide a wider investigation of a single issue in animal cognition research, namely on how researchers report and interpret non-significant findings and is the final Chapter of secondary data analysis in this thesis.

7. Chapter 7: Reporting and interpreting statistically non-significant results in animal cognition research

Null hypothesis significance testing (NHST) is the primary method of statistical analysis in animal cognition research, although the lack of a reference for this truism highlights the scarcity of descriptive research in the areas of this thesis in animal cognition. However, when NHST produces results that are not statistically significant, these are often difficult to interpret. If researchers test null hypotheses (i.e., there are no differences between a group or conditions), a non-significant result could result from a lack of any effect in the population (i.e., a true negative), or a failure to detect some true difference (i.e., a false negative). Researchers who design studies with high statistical power to detect effect sizes of theoretical interest aim to minimise the probability of these false negative errors. For example, a statistical power of 90% to detect a given effect size means that, in the long run, the researcher would correctly reject the null hypothesis 90% of the time, if the pre-specified effect size were correct. However, how negative results are reported and interpreted following the NHST logic, has been criticised on several grounds (Gigerenzer et al., 2004; Lambdin, 2012). The most prominent criticism is that researchers often misreport or misinterpret non-significant results as showing that there was no effect in the specific sample tested, or in the population at large (Aczel et al., 2018; Fidler et al., 2006; Hoekstra et al., 2006), even when these null hypotheses might be implausible (Cohen, 1994; Gelman & Carlin, 2014). In this chapter, I explore how researchers have reported and interpreted non-significant results in animal cognition and related fields, using a manually extracted dataset of negative claims following NHST from over 200 articles (Study 1). In Study 2, I further examined the *p*-value distribution of these negative results to assess if they deviated from a uniform distribution and various simulated distributions, and extended this study of p-value distributions to a larger and automatically extracted dataset of animal cognition results (Hartgerink, 2016).

7.1. Null hypothesis significance testing and *p*-values

When using NHST, researchers attempt to reject a statistical model (the null hypothesis) with their data while controlling the rate at which they will make false-positive decisions in the long-term (Neyman & Pearson, 1933). Most often, this statistical null is that there is absolutely no difference between two groups or conditions (for example a mean difference of 0 for a *t*-test; 'nil' hypothesis; Cohen, 1994), or in the case of a one-tailed test, that the difference will not be zero *or* that it will not be in a certain direction, i.e., researchers make a directional prediction for their alternative hypothesis. A statistical test then

produces a *p*-value, i.e., the probability of observing the researcher's data or more extreme data if the null hypothesis and all its assumptions were true, $Pr(d(X) \geq d(x_0); H_0)$, shortened hereafter to $P(D|H)$. If the *p*-value is lower than a pre-specified threshold (the α level), the statistical null hypothesis is rejected in favour of an alternative hypothesis (Neyman & Pearson, 1933), whereas if the *p*-value is larger than the pre-specified threshold, the statistical null hypothesis should not be rejected. However, how researchers should behave towards their null and alternative hypotheses following a non-significant result has been a continued locus of criticism of NHST in science.

When performing NHST, researchers can make statements about the long-run error probabilities of their test procedures. For example, with an *α* level of .05 and if no *α*-inflating research practices were used (Simmons et al. 2011), they can say that in the long run they would not reject H0 more than 5% of the time, if H0 were true. Similarly, if the design had 90% power to detect their smallest effect size of interest, in the long run they would only fail to reject H_0 10% of the time, if the smallest effect size did exist in the population.

7.2. Accepting the null: How much of an error?

Formally, it is an error to conclude that there is evidence in favour of the null following a non-significant result. The arbitrary nature of the *α* level highlights this: say we calculate a *p*-value of 0.08 with an *α* level of .05. By not rejecting H_0 in this instance, we can say that in the long run we would not reject H_0 more than 5% of the time, if it were true, when performing this procedure. However, if we had chosen an *α* level of .10 instead, we should have rejected H_0. Clearly, then, the p-value when using NHST is not a direct indication of the strength of evidence for or against H_0, but must be interpreted relative to error rates. However, despite the *p*-value not being the probability of the null hypothesis being true, survey studies suggest researchers do interpret *p*-values in such a way (e.g. Goodman, 2008). Moreover, scientists often misreport non-significant results as evidence of absence of a difference between groups of conditions or evidence of no effect when this inference is not necessarily warranted. For example, Hoekstra et al., (2006) reported that 41% of articles containing non-significant results in 1994-5 in *Psychonomic Bulletin & Review* interpreted the non-significant results as "evidence of no effect", a figure which rose to 60% in 2002-4. Similarly, Fidler et al. (2006) found that 63% of articles in 2000-1 in *Conservation Biology* and *Biological Conservation* reported non-significant findings as "evidence of no effect". More recently, Aczel et al., (2018) found that 72% of non-significant results were reported as "no effect" in the abstracts of 2015 articles in *Psychonomic Bulletin & Review*, *Journal of Experimental Psychology: General*, and *Psychological Science.* Such an error might be especially important in animal

cognition research, in which a combination of small sample sizes and low trial number may limit the ability of researchers to design studies with high statistical power to detect the minimum effect size of theoretical interest, and who also sometimes ask all or nothing questions about whether certain animals have certain cognitive abilities or not.

While accepting the null is an error, just how severe an error it is requires discussing. In their survey of 86 *Psychonomic Bulletin & Review* Hoekstra et al. (2006, p. 1036) reported that: "We found the *serious* mistake of accepting the null hypothesis and claiming no effect in 60% (CI: 53%, 66%) of the articles that reported statistically nonsignificant results" (emphasis added). And interpreting a non-significant result as if there were no differences between conditions ranks at number 2 of Goodman's (2008) "Dirty Dozen" *p*-value misconceptions. However, just because a researcher might report the results of significance tests incorrectly, this does not mean that they themselves, or their readers, necessarily interpreted the significance test incorrectly. In their 1933 paper, Neyman-Pearson often talked about accepting H0 following a result that was not statistically significant (Neyman & Pearson, 1933). In fact, as Mayo (2018, p. 135) writes, Neyman used the term acceptance as shorthand, and even preferred the phrase "No evidence against [the null hypothesis] is found" to "Do not reject [the null hypothesis]" (Neyman, 1976, postscript, p. 749). If scientists equate phrases such as "there were no differences between conditions ($p > 0.05$)" or "therefore we accept H_0" with "there was no statistically significant difference between the conditions" or "therefore we do no reject H_0", then the "serious mistake" of accepting the null becomes an issue of precision in language, rather than an egregious error. This is exemplified in cases where the observed experimental data are clearly more in-line with the null hypothesis than the experimenters' hypothesis of interest. For example, consider a researcher comparing10,000 wild and 10,000 hand reared birds' latencies to approach a novel object (with 99% power to detect a pre-specified effect size of interest of 2 seconds), and observed a difference of 0.02 seconds. This difference may even be statistically significant, but the effect size is clearly more compatible with the null hypothesis of 0 difference than the researchers minimum effect size of interest. Here, saying "there was no difference between the latency of wild and hand reared birds to approach the novel object ($p > .05$)", although literally incorrect, does not seem to be an error of great consequence.

7.3. Exploring non-significant result reporting and interpretation in animal cognition

Understanding how animal cognition researchers have reported and interpreted non-significant findings is an important step to, i) identify how often negative conclusions in animal cognition might be a result of NHST misreporting or misinterpretation, and ii) highlighting areas in which animal cognition

researchers can improve their statistical inferences and reporting to maximise the information they extract from negative results, and communicate this clearly. In this Chapter, I designed and led a project which explored how researchers in fields related to animal cognition report and interpret non-significant results, building on the methods used in similar studies in psychology and conservation biology (Study 1; Aczel et al., 2018; Fidler et al., 2006; Hoekstra et al., 2006). During this project, we were also able to extract the p-values associated with the negative results in our sample, which provided an opportunity to examine the distribution of this p-value and compare it to various hypothetical distributions. This comparison was performed in Study 2, and builds a picture of the properties of studies that are producing negative results in animal cognition research. To complement this analysis, I made use of a larger, automatically extracted dataset of animal cognition *p*-values (Hartgerink, 2016), which also enabled data on publication bias to be assessed.

7.4. Study 1: Reporting and interpreting non-significant results in animal cognition

In order to investigate how animal cognition researchers report negative results, we (the project team) manually extracted text and data of non-significant results from 18 journals in animal cognition, behaviour and welfare, one pre-print server, and from articles recommended by PCI: Animal Science in Study 1. We extracted data from articles reporting non-significant findings in their titles, abstracts and results section and classified how the authors interpreted them. Our classification was descriptive and aimed to characterise the different ways in which researchers reported the results of non-significant findings, and how these were translated into claims about theories and populations. Specifically, for negative results in the abstract or results section about the specific sample tested in the study, we classified the negative result text into 3 categories: 1) “Non-significant or literally correct” interpretations that either reported that there was no *significant* difference between two conditions, or words to that effect, or reported a correct directional statement; 2) “No effect” interpretations that stated there was no a difference within the sample, when in fact there was — it was just not significant in the analysis; 3) “Ambiguous” interpretations that were statements about the results that neither suggest that samples were the same, nor that there was no significant difference. Similarly, for claims about the population in the abstracts or titles, I had three related categories: : 1) “Justified”: A statement that commented on statistical power, uses equivalence tests or otherwise justifies why a non-significant result suggests that there is no theoretically important difference in the population, or that the study provides no strong evidence of a difference, 2) “Caveated, ambiguous or similar”: An interpretation of the non-significant results as suggesting/indicating etc. that X and Y do not differ in the population, or showing that they are

similar or 3) “No effect”: An interpretation of the non-significant result as showing that X and Y do not differ in the population.

7.4.1.Methods

7.4.1.1. Sample

We extracted data from a total of 20 sources, comprising 18 peer-reviewed journals, one pre-print server, and articles recommended through Peer Communities In. The 20 sources are detailed in Table 13.

Table 13: Sources of articles containing negative results in their abstracts from outlets searched back in time from March 2021.

Source	N articles
Animal Behaviour	13
Animal Behavior and Cognition	14
Animal Cognition	17
Animals	15
Applied Animal Behaviour Science	15
Behaviour	14
Behavioural Processes	15
Ethology	16
Frontiers in Psychology: Comparative Psychology	14
Frontiers in Veterinary Science: Animal Behaviour and Welfare	15
International Journal of Comparative Psychology	13
Journal of Applied Animal Welfare Science	15
Journal of Comparative Psychology	15
Journal of Ethology	15
Journal of Experimental Psychology: Animal Learning and Cognition	16
Journal of Zoo and Aquarium Research	15

Learning and Behavior	15
PeerJ: Animal Behaviour	15
bioRxiv: Animal Behaviour and Cognition	14
PCI: Animal Science	2

7.4.2. Data extraction and Classification

Myself, AV, KB, EGP, LoN, PL, SF, EL, ME and LO[12] performed the coding and were each assigned 2 journals. Coders worked back through the sources from the most recent articles available in March 2021, until they had identified 15 articles containing negative statements in the titles or abstracts that corresponded to non-significant results from null-hypothesis significance tests in the article, or until all articles in that source had been viewed. If coders were uncertain about whether an article should be included, they continued until they had 15 that they were confident with, explaining why three journals had 16 or 17 articles extracted. Each coder screened the abstracts of each article of their assigned journals and identified any negative statements about either i) the specific sample tested in that study or ii) a wider population. If a negative statement was present, the coder then recorded the paper's information (title, first author, journal and year) and the negative statement. For papers with multiple negative statements for either the sample or the population, the coder recorded the negative statement that they thought was most clearly related to the paper's main claim, such that each paper had a maximum of one negative sample statement and one negative population statement. Next, the coder verified that the paper reported the results of NHST. If verified, the coder then extracted the text of the NHST that corresponded to the abstract claim from the results section of the manuscript, including the associated *p*-value. If there was more than one corresponding statistical test within an experiment, coders extracted the test result that they thought was most relevant to the claim. If the abstract claim was equally supported by multiple experiments, coders extracted the information from the first experiment presented.

After the title, abstract claims (sample and population), result text and *p*-value had been extracted, the coder categorised how each negative statement was reported. Through piloting, discussion, from looking at the previous studies (Aczel et al., 2018; Fidler et al., 2006; Hoekstra et al., 2006), and discussion with a previous author (Aczel, personal communication), I developed three categories. For the

[12] AV: Alizée Vernouillet; KB: Katharina Brecht; EGP: Elias Garcia-Pelegrin; LoN: Laurie O'Neill; PL: Poppy Lambert; SF: Shannon Francis; EL; Ed Legg; ME: Mahmoud Elsherif; LO: Ljerka Ostojić

sample claims and result text, these were: 1) “Non-significant or literally correct” interpretations that either reported that there was no *significant* difference between two conditions, or words to that effect, or reported a correct directional statement; 2) “No effect” interpretations that stated there was not a difference within the sample, when in fact there was — it was just not significant in the analysis; 3) “Ambiguous, Similar or small effect size” interpretations that were statements about the results that neither suggest that samples were the same, nor that there was no significant difference between them (which were later split into “Ambiguous” and “similar or small effect size”). In addition to these descriptions, I developed a table of hypothetical statements that are detailed in Table 14, that were available to the coders during the project.

Similarly, the title, if it contained a negative statement, and population claims from the abstracts were categorised into three categories: 1) “Justified”: A statement that commented on statistical power, use of equivalence tests or otherwise a justification why a non-significant result suggests that there is no theoretically important difference in the population, or that the study provides no strong evidence of a difference, 2) “Justified, caveated or ambiguous”: An interpretation of the non-significant results as suggesting/indicating etc. that X and Y do not differ in the population, or showing that they are similar, and 3) “No effect”: An interpretation of the non-significant result as showing that X and Y do not differ in the population. In addition to these descriptions, I developed a table of hypothetical statements that are detailed in Table 15. The exact coding manual used is available at https://osf.io/gdp6f/.

Table 14: Example categorisation of sample-level claims

Category	Non-significant or literally correct	No Effect	Ambiguous, Similar, or small effect size
Description	Reports that there was no *significant* difference between two conditions, or words to that effect.	A statement that there was not a difference within the sample, when in fact there was – it was just not significant in their analysis.	A statement about the results that neither suggests they were the same, nor that there was no significant difference.

Examples	There was no significant/detectable difference between X and Y. We did not detect a difference between X and Y (or any other statement implying failing to find a signal within noise). We did not find a significant effect. X was not significantly related to Y. X did not perform significantly above chance. X performed significantly above chance, but Y did not. There were no significant differences between X and Y's performance. X did not do A more in condition Y than condition Z (and this is genuinely true in the data – see Note).	There was no difference between X and Y. There was no effect. There was no evidence of an effect. There was no relationship between X and Y. We did not find/observe/see a difference between X and Y. We did not find an effect. We found no evidence of an effect. X performed at chance levels. X performed above chance, but Y did not (if Y > chance, but not significantly). X and Y performed equally. X did not do A more in condition Y than condition Z (but this is not true in the data). We did not find/observe/see a relationship between X and Y. We did not observe X performing above chance.	X and Y were similar. There was no large/clear difference between X and Y. There was no large effect of X on Y

		We found that X performed above chance, but Y did not.	

.

Table 15: Example categorisation of population-level or title claims

Category	Justified	Caveated, Ambiguous or Similar	No Effect
Description	Comments on statistical power, uses equivalence tests or otherwise justifies why a non-significant result suggests that there is no theoretically important difference in the population, or that the study provides no strong evidence of a difference.	Interprets the non-significant results as suggesting/indicating etc. that X and Y do not differ in the population, or are similar.	Interprets the non-significant result as showing that X and Y do not differ in the population.
Examples	Because the test was high-powered to detect a meaningful difference, this non-significant result suggests that A is not related to Y in a theoretically important way. In addition to being not statistically different to each	…Suggesting that X is not related to Y. …Indicating that X is not related to Y. …Suggesting/indicating that there is no difference between X and Y.	… Meaning that X is not related to Y. … Showing that X is not related to Y. There is no difference between X and Y. X and Y do not differ. X and Y are similar. X and Y are the same (show the same effect, etc).

	other, X and Y were also statistically equivalent (if a frequentist equivalence or non-inferiority test was performed), suggesting that X is not meaningfully related to Y. Any statement about the test likely being low powered and this making it difficult to interpret what the results mean at the population level.	Suggesting that X has not changed Y. Our results provide no strong evidence that X and Y are different. Suggesting that X and Y are similar.	X does not change Y. Our results provide no evidence that X and Y are different.

7.4.3. Reliability and Quality Control

Twenty-four articles (8.5%) were double-blind coded in order to assess the likely reliability of our coding scheme, and all articles underwent a quality control procedure involving a second coder to identify any mistakes or inconsistencies.

7.4.3.1. Double blind extraction

BGF independently coded 24 articles, namely the first four articles from six randomly chosen journals, blind to the results of the original coders. From this, I computed inter-rate agreement for each variable that was extracted (Title Population Claim Level; Title Sample Claim Level; Abstract Sample Claim Text; Abstract Sample Claim Level; Abstract Population Claim Text; Abstract Population Claim Level; Result Text; Result Level; *p*-value).

7.4.3.2. Quality control

All articles underwent quality control. Here, a second coder reviewed the data extracted from each article. The quality controller verified 1) whether a negative claim from the title/abstract have been

extracted, 2) that any negative claim extracted was really a negative claim, 3) whether the result that was extracted corresponded to the claim that was extracted, and 4) whether they agreed with the classification of each claim. If the quality controller identified a mistake, they classified this as a major disagreement, whereas if the quality controller disagreed but was uncertain about this judgment, for example in the case of borderline claims, they classified this as a minor disagreement. BGF reviewed all disagreements and made a final decision on what entered the final dataset, returning to the original article if necessary.

7.4.4. Analysis

My primary analysis was descriptive. I present the percentage of claims in each category across the titles, abstract population claims, abstract sample claims, and result texts. To illustrate the types of claims placed in each category, I provide examples in tables. In addition, every categorisation can also be viewed in the open dataset. I used a Chi-squared test to compare whether, if a "no effect" interpretation was made in the results, it was more likely that a "no effect" interpretation would also be made in the abstract, compared to when a correct interpretation was made in the results

7.5. Study 1 Results

I extracted data from 302 articles. Of these, 18 were excluded due to their identified claim having no corresponding negative result of a NHST (e.g., if only descriptive statistics were used, or only a Bayesian analysis performed) and one was excluded due to excessive ambiguity in how the results were described. This left a final sample of 283 articles for analysis.

7.5.1. Reliability and Quality Control

7.5.1.1. Double Blind Coding

For 24 articles (8.5% of the total sample), two authors (BGF and the author originally assigned to the journal) extracted all the data independently of each other. Only 5 of the article titles were identified as containing negative statements by either of the two coders, and from this, the two coders agreed on only 1 out of 5 (20%) of the articles about whether the title statement was about the tested sample or the population. Following discussion with the whole group, we agreed that it was often ambiguous whether the titles of articles were referring to the specific sample tested or a wider population, and so we decided to combine these measures and have no sub-group analysis for the title claim, deviating from our original plan. When considering the category (justified; caveated, similar or ambiguous; no effect) of the title claim, the two coders agreed on two out of six papers (33%). Three of the four disagreements occurred

when one coder did not interpret the title as a negative claim, e.g. as in “Evidence that novel flavors unconditionally suppress weight gain in the absence of flavor-calorie associations” (Seitz et al., 2020), and one where a coder appeared to have made an error. From discussion within the group, it was evident that these ambiguous cases — where the statements were not clearly written as negative statistical results, but more described some theoretical conclusion — proved the largest source of difficulty during the whole coding procedure, and this affected the reliability of the title claims and population claims from the abstract.

The coders identified 24 sample claims from the abstracts of the papers, from which they coded the same claim on 22 out of 24 occasions (91.6%). Of these 22 claims, the two coders agreed on 19 of their levels (86.3%). In contrast, the coders identified only eight population claims from the abstracts of articles, from which they agreed on three occasions (37.5%), and of these three, agreed on two of their levels (66.7%). From the results, coders recorded the same result for 16 of the 22 (72.7%) abstract claims they coded the same, and of these 16, agreed on 13 of their levels (81.3%), and extracted the exact same *p*-value for 10 of these 13 (76.9%).

In sum, the double-blind coding demonstrated good inter-rater consistency for how the abstract sample claims and associated results and *p*-values were extracted, even before our quality control procedures had been implemented. In contrast, inter-rater consistency was low for the title claims and population claims from the abstracts. This matched our subjective experience of the coding procedure, where we experienced many cases of population claims as vague and about a theoretical hypothesis that did not closely correspond to any particular negative result from the article. In contrast, the negative sample claims are often easily mapped onto a particular negative result in the text.

7.5.1.2. Quality Control

Each article was checked by a quality controller. The initial coders identified 67 possible negative statements in the titles of papers, and the quality controller agreed with the classification of 39 (58%) of these statements, had a minor disagreement with six statements (9%), and a major disagreement with 22 statements (33%). Of note, 16 of these 22 major disagreements came from a single repeated error in which an individual coder coded ‘ambiguous’ for titles containing no negative statement. In the abstract, coders identified 281 negative statements about the specific sample tested in the paper. Of these, the quality controllers agreed with the classification of 250 (89%), had minor comments about 16 (6%), and major disagreements with 15 (5%). Coders identified a much smaller number of negative inferences

about populations in the articles and disagreed more frequently: Of the 82 identified statements, the quality controllers agreed with the classification of 44 (53%), had minor comments about 18 (22%) and major disagreements with 20 (24%). Regarding the result texts from the article bodies, coders identified 282 results, of which the quality controller agreed with the classification and extracted *p*-value for 252 (89%), had minor comments for 13 (5%), and major disagreements for 17 (6%).

The quality control process allowed us to i) identify any clear errors in the data extraction process, ii) highlight borderline cases where our coding scheme could not clearly categorise certain statements, and iii) assess the robustness of the coding procedure. In line with the results from the double-blind coding, the quality control process demonstrated a high inter-rater agreement and consistency with identifying and classifying negative sample statements from abstracts, and the corresponding results and *p*-values from the main text, yet greater inconsistency in deciding, i) whether titles and population statements were truly "negative" in the sense of being the result of a non-significant NHST, and ii) whether the authors were claiming the absence of an effect from these negative results. This inconsistency occurred mainly because many titles and population statements referred not to a certain statistical result but made a vague theoretical statement. An example of the former would be 'Dogs did not bark more in condition A than condition B', whereas an example of the latter would be 'Our results suggest that dogs display similar levels of anxiety when faced with unfamiliar conspecifics and heterospecifics', where anxiety was not clearly defined in the article.

7.5.2.Title Claims

Forty-four titles contained negative statements resulting from non-significant results of null-hypothesis significance tests. Of these, 37 (84%) interpreted the non-significant result as evidence of no effect, whereas seven (16%) made caveated claims or claims about two groups or conditions being 'similar'. Table 16 provides examples of these claims.

Table 16: Examples of “No Effect” and “Caveated or Similar” claims in the titles of papers following non-significant NHST.

No Effect
N = 37 (84%)
“Home range use in the West Australian seahorse Hippocampus subelongatus is influenced by sex and partner’s home range but not by body size or paired status” Kvarnemo et al., 2021
“Delays to food-predictive stimuli do not affect suboptimal choice in rats.” Cunningham & Shahan, 2020
“Common Marmosets (*Callithrix jacchus*) Evaluate Third-Party Social Interactions of Human Actors But Japanese Monkeys (*Macaca fuscata*) Do Not” Kawai et al., 2019
Caveated, Ambiguous, or Similar
N = 7 (16%)
“Limited Evidence of Number-Space Mapping in Rhesus Monkeys (Macaca mulatta) and Capuchin Monkeys (*Sapajus apella*)” Beran et al., 2019
“Little Difference in Milk Fatty Acid and Terpene Composition Among Three Contrasting Dairy Breeds When Grazing a Biodiverse Mountain Pasture” Koczura et al., 2021
“The Equipment Used in the SF6 Technique to Estimate Methane Emissions Has No Major Effect on Dairy Cow Behavior” Pereira et al., 2021

7.5.3.Abstract Claims

7.5.3.1. Abstract Sample Claims

We extracted 278 negative claims about a sample result. Of these, 174 (63%) claimed evidence of no effect, 71 (26%) made formally correct statements that there were no statistically significant differences between groups or conditions, 17 (6%) made claims about an effect being ‘similar’ between groups or conditions, or described a small effect size, and 16 (6%) were ambiguous. Table 17 provides examples of these claims.

Table 17: Examples of “No Effect”, “Similar or small effect size”, “Non-Significant” or “Ambiguous” claims about the sample in the abstracts of papers following non-significant NHST.

No Effect
N = 174, 63%
“Levels of individuals sitting with their back to the window was unaffected by visitor number or noise.” Hashmi & Sullivan, 2020
“The groups did not differ in their ability to follow human signals” Lazarowski et al., 2020
Similar or small effect size
N = 17, 6%
“Pair members demonstrated comparable responses towards a male ‘intruder’, as latency to respond and proximity scores were very similar between pair members in the majority of pairs examined” DeVries et al., 2020
“We found that individuals called back to sympatric and allopatric calls within similar amounts of time,” Wu et al., 2021
Non-significant
N = 71, 26%
“Nutcrackers… did not significantly change their caching behaviour when observed by a pinyon jay.” Vernouillet et al., 2021
“No significant correlations between degree of laterality and behavioral interest in the stimuli were found” Lilley et al., 2020
Ambiguous
N = 16 (6%)
“We also found no conclusive evidence that either the visual or the vibratory sensory modalities are critical for prey capture.” Meza et al., 2021
“No systematic variations on space allocation were observed in neither experiment” Ribes-Iñesta et al., 2020

7.5.3.2. Abstract Population Claims

We extracted 63 negative claims about a population that followed on from the negative result within a sample. Of these, 45 (71%) were caveated and 18 claimed that there was no effect (29%). Table 18 provides examples of these claims.

Table 18: Examples of "No Effect" and "Caveated, Ambiguous or Similar" claims about populations in the abstracts of papers following non-significant NHST.

No Effect
N = 18 (29%)
"Partial rewarding does not improve training efficacy" Cimarelli et al., 2021
"Our findings show that H. *horridum* does not respond to hypoxic environments" Guadarrama et al., 2020
"Oviposition site choice is not by-product of escape response" Kawaguchi & Kuriwada, 2020
Caveated, Ambiguous, or Similar
N = 45 (71%)
"These results suggest capuchin monkeys do not engage in indirect reciprocity" Schino et al., 2021
"These results suggest that shoal composition may not be an important driver of shoal choice in this system" Paijmans et al., 2021
"…suggesting that size is not a determinant factor for feral horse society." Pinto & Hirata, 2020

7.5.4.Result Text

In the results sections, 276 non-significant results of NHST were coded. Of these, 140 (52%) were reported as not significant, 113 (41%) as if there was no effect, 12 (4%) were reported as groups or conditions being similar, 10 (4%) were coded as ambiguous, and one (0.4%) reported a "trend" in the opposite direction to the prediction. Several of the ambiguous codes were due to authors' use of "main effect" when interpreting ANOVA terms, where we thought that saying there was "no main effect of X" was different enough to saying "no effect of X" to not be included in the "No Effect" category, although

this highlights the somewhat arbitrariness of our categories. Table 19 provides examples of the different types of result reporting.

Table 19: Examples of "No Effect", "Similar or Small Effect Size", and "Non-Significant" claims in the results sections of papers following non-significant NHST.

No Effect
N = 113 (41%)
During farrowing, No Effect of the treatments was seen on the percentage of time spent (3.22 % vs. 1.90 %, P = 0.372) on the nest-building behaviour" Aparecida Martins et al., 2021
"There were no differences between treatments in the frequency or duration of birds flying between walls" Stevens et al., 2021
Similar or small effect size
N = 12 (4%)
"The average time yaks spent grazing was similar among shrub coverage groups (P = 0.663)" Yang et al., 2021
"The number of sessions required to reach criterion didn't reliably differ between groups" O'Donoghue et al., 2020
Non-significant
N = 140 (52%)
"Comparing the pooled data of all crows, no significant increase in the number of mark-directed behaviors during the mirror mark condition was found compared with the no-mirror sham condition." Brecht et al., 2020
"There was no significant effect of removal type on changes in display strength in either dominant males or subordinate males." Piefke et al., 2021
Ambiguous
N = 10 (4%)
"As can be seen in Figure 1D, there was no difference in response rates after R and NR trials across days for rats under reward uncertainty." [where in Figure 1D the bars on the graph look almost identical) Anselme & Robinson, 2019
"It showed that there was a significant main effect of session, but no main effect of CS" Harris & Bouton, 2020

Notably, if a “No Effect” interpretation was made in the results, it was more likely that a no effect interpretation would also be made in the abstract, compared to when a non-significant interpretation was made in the results ($\chi 2(1, N = 211) = 21.65$, $p < .0001$). Limiting the data to just those with “non-significant” or “no effect” responses in the abstract and results, of the 92 “no effect” statements in the results, 80 (87%) of the corresponding sample statements were interpreted as no effect. In contrast, of the 119 non-significant statements in the results, only 67 (56%) of the corresponding sample statements were interpreted as there being no effect. Nevertheless, no effect interpretations in the abstracts were absolutely the most likely for both no effect and non-significant results statements.

7.6. Study 1 Discussion

Study 1 extracted and categorised how animal cognition researchers reported the results of non-significant null hypothesis significance tests in 253 articles between 2019 and 2021. Across titles, abstracts and results, researchers often reported non-significant results with the “no effect” phrasing that has often been labelled as erroneous (titles 84%; abstract sample results 63%; result text 41%). However, reporting these results as “not statistically significant” was also common – most often in the results section (titles 16%; abstract sample results 26%; result text 52%). The other, albeit less frequent, method of reporting non-significant results was to comment on the similarity between two groups or conditions (abstract sample results 6%; result text 4%).

Overall, these results demonstrate considerable heterogeneity in how animal cognition researchers report non-significant results, something that was also observed within categories (i.e., how researchers phrased the “No Effect” or “Non-Significant” interpretations, although this was not directly assessed). Moreover, they suggest that negative results are at risk of misreporting and misinterpretation in animal cognition publications. It remains a question, however, what the consequences of such misreporting might be, *i.e.*, how readers of scientific articles interpret “No Effect” statements, and this could be studied through analysing how these studies are cited, in other publications but also in media reports and student essays. It is likely that many scientists write “No Effect” with the assumption that it would be read as “No Significant Effect”, and even if they did not, readers of articles may interpret “No Effect” as “No Significant Effect” anyway.

Possibly encouragingly, when researchers extended no effect statements from the sample to the population, they routinely opted for qualifies to caveat inference to the populations (e.g., …these results

suggest that there is no effect at the population level). Again, however, more research is needed to understand how such statements are interpreted and implemented by scientists and the wider community. The caveating of population statements likely reflects the lack of a formal strategy being used to interpret negative results, such as equivalence tests or Bayes factors. Although beyond the scope of the current Chapter, Lakens (2017) provides a detailed tutorial for equivalence testing in psychological research, and Rose et al., (2018) in animal behaviour, and Rouder et al. (2009) provide an introduction to Bayes Factors.

Notably, my coding team found it difficult to identify and classify negative population statements in the abstracts of articles. This likely reflects the distance between the theoretical claims researchers wish to test and the actual statistical hypotheses that are tested, *i.e.,* rarely can a theoretical prediction about an animal's cognition be reduced to a single decisions between a null and alternative hypothesis in a null hypothesis significance test.

Finally, we found that "No Effect" interpretations were more common in abstracts and titles than they were in the result text. That is, authors who have written out "non-significant" interpretations in the results nevertheless wrote "no effect" interpretations in the abstracts and titles. This could be due to two factors: word limits and incentives to make bolder claims, the former of which should be considered by journal editorial boards when setting policy.

Next, in Study 2, I furthered the exploration of negative results in animal cognition by generating the *p*-value distribution of all the negative results that we extracted as part of Study 1, and complemented this with an analysis of an open database of text-mined NHST results (Hartgerink, 2016)

7.7. Study 2: *p*-value distributions of manually and automatically extracted negative results in animal cognition

During the data-extraction process of Study 1, I extracted the *p*-values corresponding to the negative statements in animal cognition paper's abstracts. The distribution of these *p*-values can provide information on how often and how extreme the evidence against null hypotheses is produced across the cohort of studies that data was extracted. For example, if all studies tested against null hypotheses that were really true, the *p*-value distribution would follow a uniform distribution (Simonsohn et al., 2014). In contrast, if all studies tested against a null hypothesis that was false, the p-value distribution would exponentially decay, the rate of which would be determined by the power of the studies to detect the true effect size. To examine this, I therefore compared our observed *p*-value distribution from the

manually extracted data to five other distributions. First, I compared the distributions to a uniform distribution – the distribution that would be expected if the null hypothesis was true across every study. Next, I compared our observed distribution to four simulated distributions. These simulated distributions consisted of research where the alternative hypothesis was correct 80% of the time, and studies had either 10%, 33%, 50% or 80% power to detect the simulated effect size.

To further this analysis, I used an open dataset of NHST results by Hartgerink (2016) that were machine-extracted by StatCheck. StatCheck is an R package (Nuijten & Epskamp, 2015) with a web interface (http://statcheck.io/) that extracts and analyses APA-style conforming results from NHST tests. Most often, StatCheck has been used to check the consistency between p-values reported in papers, and those automatically re-calculated by StatCheck. StatCheck achieves this by computing p-values from the machine-identified and machine-read test statistics and degrees of freedom (Nuijten et al., 2016; Nuijten & Polanin, 2020). However, pertinent to the current paper, Hartgerink (2016) openly archived 688,112 NHST results that had been extracted and analysed by StatCheck from 322 psychology journals. Three of these 322 journals were animal cognition journals, and ones that we had manually extracted data from in Study 1: *Animal Cognition*, *Learning and Behavior*, and *Journal of Comparative Psychology*. In total, Hartgerink archived 14,217 results from these three journals, containing 4,758 negative results, which I explored in this study. I explored three features of the StatCheck-extracted data, which were as follows.

First, and similar to the manually extracted dataset, I compared the *p*-value distribution of all non-significant results to that expected by i) a uniform distribution where H_0 was true for 100% of studies, and ii) four simulated distributions with H1 true for 80% of studies, which had either 10, 33, 50 or 80% power to detect the simulated effect size. Second, as the StatCheck dataset also included significant *p*-values, I examined the distribution of p-values around the conventional significance level of .05, as a possible indicator of publication bias. I again used the simulated distributions as a comparison, and a drop in published p-values just above .05 would indicate some form of publication bias – either not reporting some non-significant *p*-values, or at least not highlighting them in text. Finally, I examined the frequency of StatCheck identified "decision errors" across the three animal cognition journals. StatCheck labels a result a decision error if the re-calculated *p*-value is on the opposite side of .05 to the p-value reported in the article, and hence would lead to a different inference under the standard NHST paradigm with an α level of .05.

7.8. Study 2 Methods

7.8.1. Extracting animal cognition data from the StatCheck Dataset

I imported the Hartgerink (2016) dataset into R 4.0.2 and filtered the dataset by the journal column, extracting the data for “Animal Cognition”, “Journal of Comparative Psychology” and “Learning and Behavior”, using the filter function from ‘dplyr’ in the ‘tidyverse’ (Wickham et al.,2019). This dataset contained 14,217 p-values, 859 from Animal Cognition from 2003 to 2016, 11059 from Journal of Comparative Psychology from 1985 to 2016, and 2299 from Learning and Behavior between 2010 and 2015. Of these 14,217 p-values, 4,758 (34%) were calculated by StatCheck as > .05.

7.8.2.p-value distributions: manually extracted data

For the *p*-value analysis, I plotted the distribution of p-values and used a two-sided Kolmogrov-Smirnov test to compare this distribution to the uniform distribution that would be expected if H0 was correct in every case. The two-sided Kolmogrov-Smirnov test tests whether the cumulative density function (CDF) of a distribution is equal to that of another, in this case the observed distribution and a uniform distribution across the interval .05-1.0. Because *p*-values in the interval .05-0.10 are often taken as “trend” or “marginal” evidence, and thus may be under considerably different publication biases to *p*-values in the interval .10-1.0, I performed a second Kolmogrov-Smirnov test comparing our observed distribution with a uniform distribution in the interval 0.10-1.0. Next, I compared the observed distribution to four simulated distributions. These simulated distributions were mixture distributions that assumed that a body of research was conducted where 80% of the time the alternative hypothesis was correct and studied with either 10%, 33%, 50% or 80% power to detect the simulated effect size. The simulations were performed by taking two samples from two normal distributions, with a different mean when H1 was correct, and the same standard deviation. These were compared using a two-sided *t*-test, and the *p*-values collected across all simulations. Ten thousand simulations were performed to construct each distribution, and then the distributions were compared qualitatively, by looking at the proportion of *p*-values across bins of .05.

7.8.3.p-value distributions: StatCheck data

For the manually extracted data, I plotted the *p*-value distribution of all non-significant statistical test results, defined as *p*-values in the interval $.05 < p <= 1$, using StatCheck’s calculated *p*-values, and compared this to a uniform distribution using a two sided one-sample Kolmogrov-Smirnov test. However, because *p*-values less than .10 are often interpreted as evidence in favour of an effect, and therefore may be under a publication bias (with *p*-values interpreted as evidence in favour of an effect more likely to be published than those that are not interpreted as supporting an effect), I also plotted the *p*-value

distribution for the interval .10 < *p* <= 1, and again compared this to a uniform distribution. In addition to the Kolmogrov-Smirnov test, I compared the observed distributions to three other simulated *p*-value distributions across a body of research where 80% of the time H1 was true, but with either 33%, 50% or 80% power. The simulations were performed as in study one, but with a new seed, and ten thousand simulations. Second, I also examined how the distribution of calculated *p*-values changed across the significance threshold. To do this, I visualised the *p*-value distribution from .01 to .15, and compared the ratio of *p*-values in each bin of .01, focusing on the ratio of *p*-values in the interval .04 < *p* < .05, and the interval .05 < *p* < .06.

7.8.4.Significance threshold and publication bias

Finally, I visualised how the *p*-value distributions changed shape around the significance threshold, which I assumed to be .05 for all papers. Specifically, as an indicator of publication bias, I compared the ratio of *p*-values in bins of .01 either side of the significant threshold (see e.g., Lakens, 2015).

7.9. Study 2 Results

7.9.1.*p*-value distributions: manual data

In total, 202 of the 283 papers reported exact *p*-values, with the other 81 reporting either inequalities or not reporting the *p*-values at all. Of these 202 *p*-values, four were below .05 and non-significant due to a lower α level. The distribution of the 198 non-significant *p*-values in the interval .05–1 is displayed in Figure 23. This distribution significantly differs from a uniform distribution (two-sided Kolmogrov-Smirnov test, $D = 0.12$, $p = .0087$).

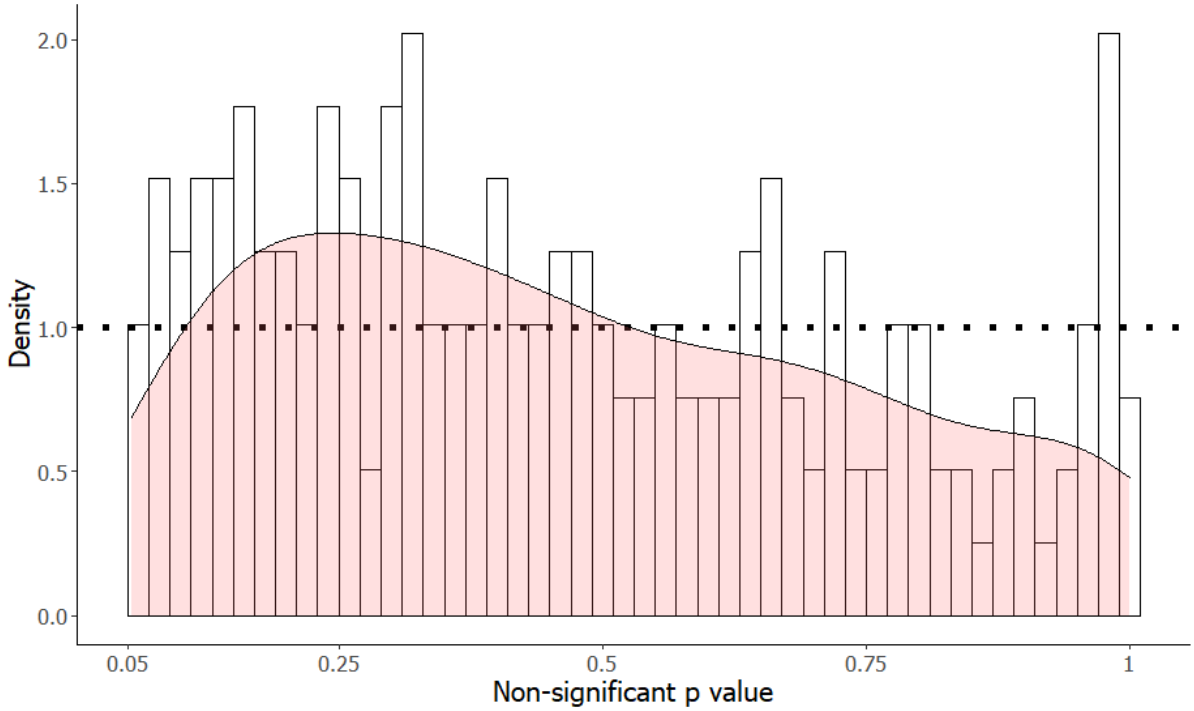


Figure 23: Histogram showing the distribution of non-significant p-values from result sections of 198 articles in animal cognition and related fields, with a density distribution overlaid in pink. The dotted line shows the average density.

Figure 24 contrasts Figure 23 with the four simulated distributions of bodies of research performed where 80% of alternative hypotheses were correct, and studies had either 10, 33, 50 or 80% power to detect the true effect size of H1 when it was true. Notably, *p*-values in the interval from .05 to .10 were underrepresented in the manually extracted data, making up only 5.6% of observations compared to 8.2% (10% power simulation), 15% (33% power simulation), 19% (50% power simulation), and 20% (80% power simulation). Similarly, very high *p*-values (.95-1.0) were overrepresented in our manual dataset (7.6% of observations, compared to 4.3%, 3.2%, 2.4% and 3.4% for the 10, 33, 50 and 80% power simulations respectively), which likely reflects either the use of multiple correction procedures, or small sample non-parametric statistics that produce non-uniform distributions under the null hypothesis.

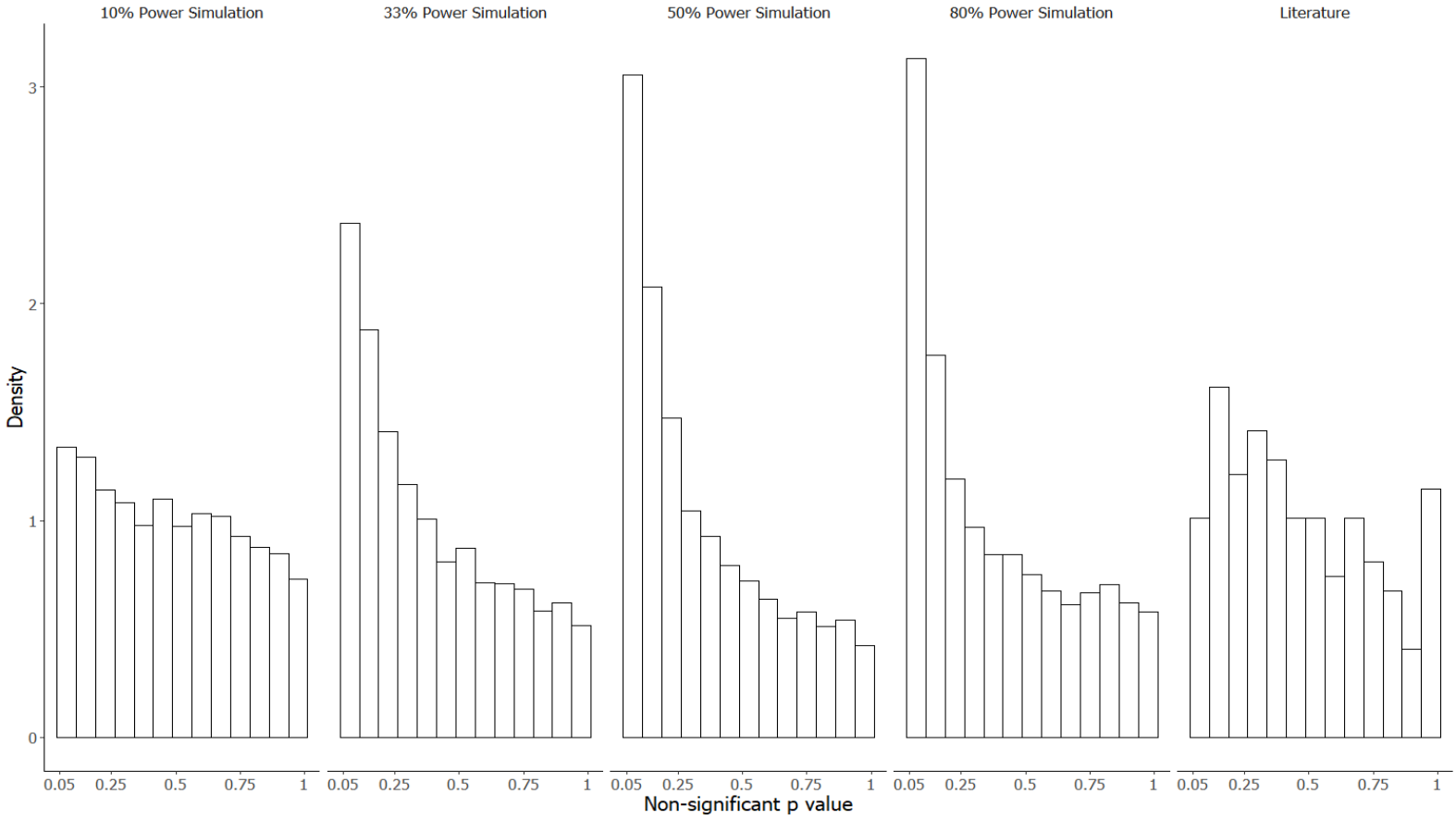


Figure 24: The observed p-value distribution of 198 p-values > .05, which were manually extracted from results corresponding to negative claims present in the abstracts of animal cognition articles, compared to 3 simulated distributions where 80% of alternative hypotheses were correct, with studies performed at either 10%, 33%, 50% or 80% statistical power.

7.9.2.StatCheck Data: Quality Assessment

During data exploration, I noticed numerous errors in StatCheck's calculations of one-tailed *p*-values, for example calculating a *p*-value that was two times greater, rather than two times smaller, for a one-tailed test compared to the corresponding two-tailed test. For this reason, I excluded all 417 detected one-tailed tests from all subsequently reported analyses.

7.9.3.*p*-value distributions: StatCheck data

The *p*-value distribution for all 4,577 non-significant results in the three animal cognition journals is displayed in Figure 25. The distribution was significantly different from a uniform distribution (two-sided one sample Kolmogrov-Smirnov test, $D = 0.23$, $p < .0001$), and this significant difference was also observed

when I restricted the analysis for all 3,815 non-significant results in the interval .10 < *p* <= 1 for the three animal cognition journals (two-sided one-sample Kolmogrov-Smirnov test, *D* = 0.16, *p* <.0001).

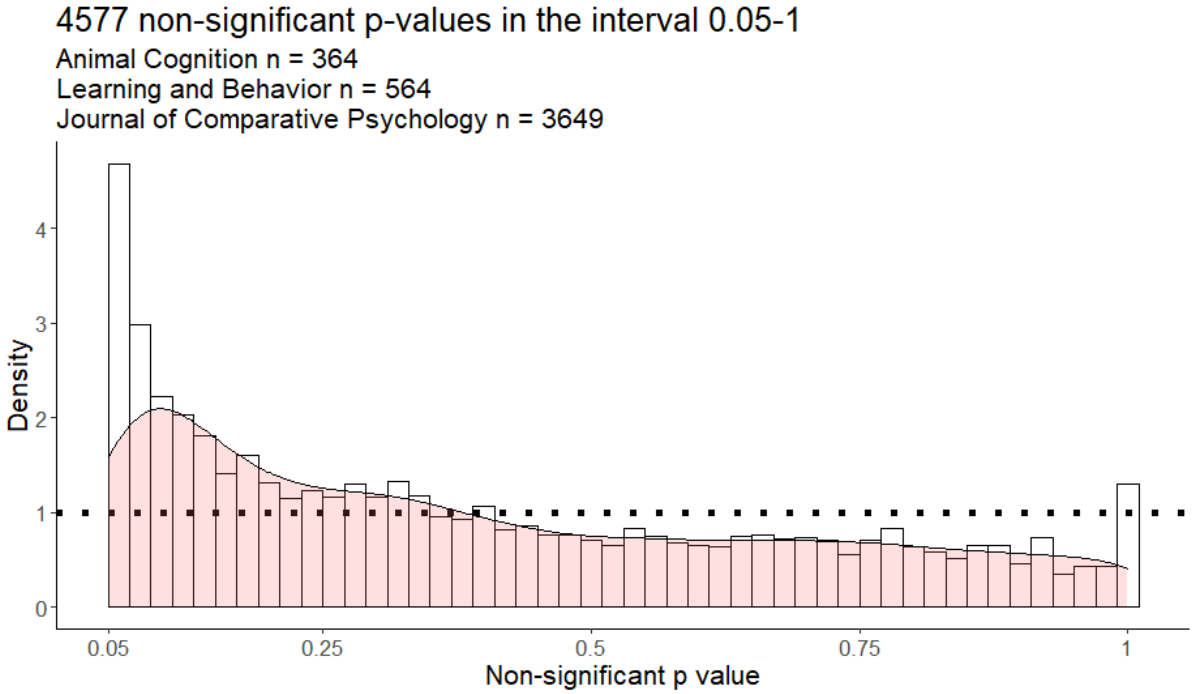


Figure 25: Histogram of the observed p-value distribution of 4577 p-values > .05 automatically extracted from *Animal Cognition, Learning and Behavior*, and *Journal of Comparative Psychology* by StatCheck, with a density distribution overlaid in pink. The dotted line plots the average density.

Figure 26 contrasts this distribution with the four simulated distributions of bodies of research performed where 80% of alternative hypotheses were correct, and studies had either 10, 33, 50 or 80% power to detect the true effect size of H1 when it was true. In contrast to our manually extracted data, *p*-values in the interval .05-.10 were not overrepresented in the automatically extracted data compared to the simulations. In the StatCheck extracted literature data, 18% of non-significant findings were in the interval .05-.10, compared to 8.0% (10% power simulation), 16% (33% power simulation), 19% (50% power simulation), and 21% (80% power simulation). However, the ratio of *p*-values in the interval .95-1.0 compared to .90-.95 was higher in the literature dataset 1.65:1, compared to the 1.02:1, 0.88:1, 1:1 and 0.81:1 observed in the 10, 33, 50 and 80% power simulations, respectively.

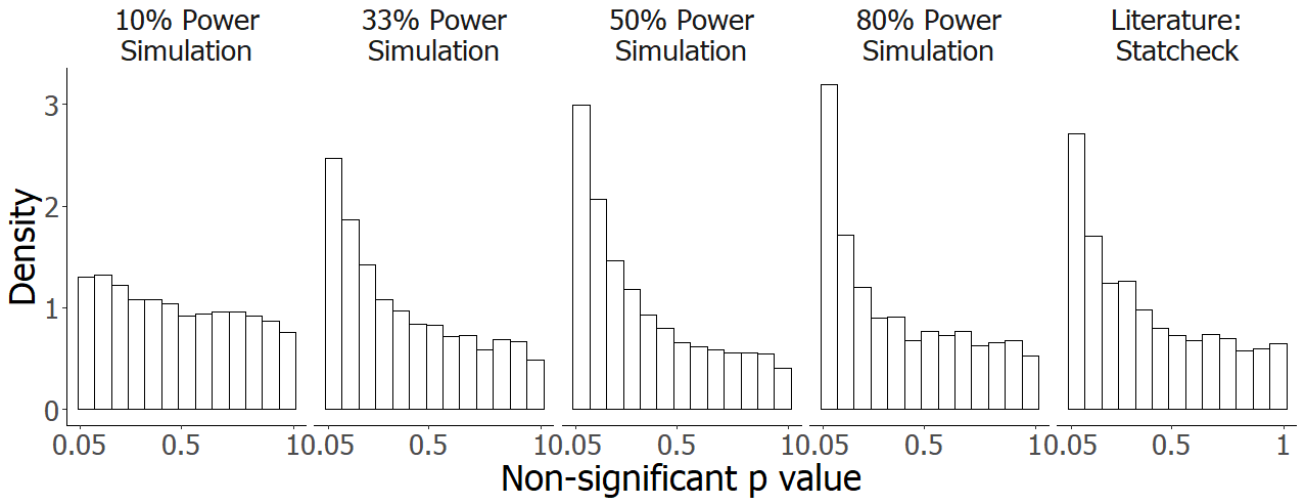


Figure 26: The p-value distribution of 4577 p-values > .05 automatically extracted from *Animal Cognition, Learning and Behavior*, and *Journal of Comparative Psychology* by StatCheck, compared to 3 simulated distributions where 80% of alternative hypotheses were correct, with studies performed at either 10%, 33%, 50% or 80% power.

7.9.4.Significance threshold

I next examined how the *p*-value distributions of the StatCheck literature data contrasted with the four simulations around the conventional significance threshold, *p* =.05. Figure 27 displays these *p*-value distributions for the interval from .01 to .15, and the corresponding binned data are presented in Table 20. Of note, the ratio of *p*-values in the interval .04-.05 and .05-.06 was higher in the literature data, 2.07:1, than in any of the simulations: 1.13:1 (10% power simulation); 1.12:1 (33% power simulation); 1.25 (50% power simulation); 1.21 (80% power simulation).

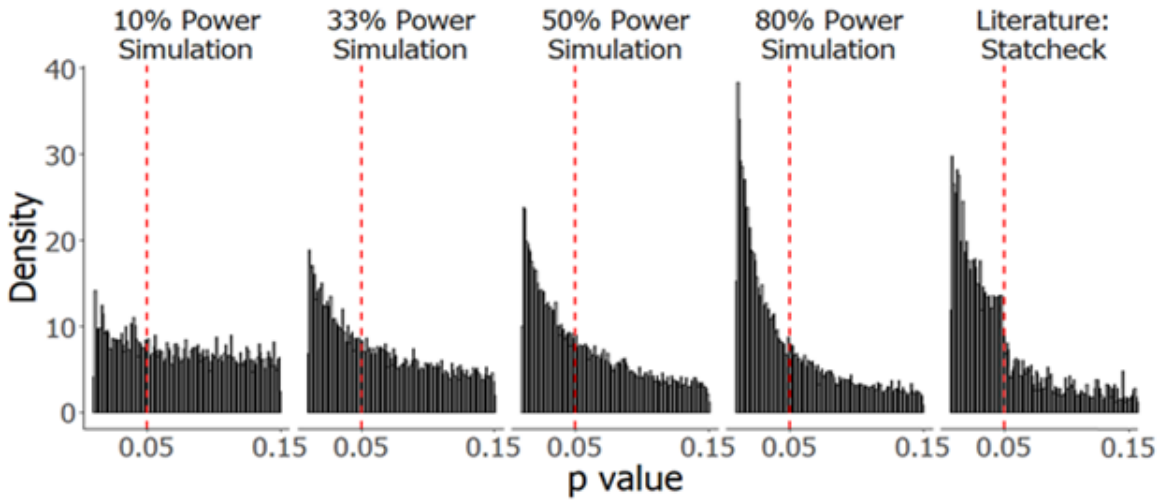


Figure 27: The p-value distribution of 3,956 p-values in the interval .01 < p < .15 automatically extracted from *Animal Cognition, Learning and Behavior*, and *Journal of Comparative Psychology* by StatCheck,

compared to four simulated distributions where 80% of alternative hypotheses were correct, with studies performed at either 10%, 33%, 50% or 80% power.

Table 20: The p-value distribution of 3,956 p-values in the interval .01 < p < .15 automatically extracted from *Animal Cognition, Learning and Behavior*, and *Journal of Comparative Psychology* by StatCheck, compared to four simulated distributions where 80% of alternative hypotheses were correct, with studies performed at either 10%, 33%, 50% or 80% power.

Interval	Literature		10% power		33% power		50% power		80% power	
	Raw	Prop	Raw	Prop	Raw	Prop	Raw	Prop	Raw	Prop
0.01-0.02	975	0.25	816	0.103	2185	0.16	3051	0.19	3571	0.27
0.02-0.03	680	0.17	665	0.084	1690	0.12	2117	0.13	2144	0.16
0.03-0.04	558	0.14	670	0.084	1353	0.096	1695	0.11	1476	0.11
0.04-0.05	492	0.12	640	0.081	1185	0.084	1397	0.089	1043	0.080
0.05-0.06	230	0.058	572	0.072	1050	0.075	1127	0.071	863	0.066
0.06-0.07	198	0.050	515	0.065	980	0.070	1045	0.066	726	0.055
0.07-0.08	140	0.035	559	0.070	855	0.061	910	0.058	596	0.045
0.08-0.09	133	0.034	530	0.067	813	0.058	888	0.056	499	0.038
0.09-0.10	108	0.027	515	0.065	740	0.053	703	0.045	473	0.036
0.10-0.11	95	0.024	518	0.065	757	0.054	656	0.042	385	0.029
0.11-0.12	94	0.024	493	0.062	650	0.046	597	0.038	357	0.027
0.12-0.13	92	0.023	485	0.061	616	0.044	564	0.036	357	0.027
0.13-0.14	90	0.023	471	0.059	631	0.045	539	0.034	322	0.025
0.14-0.15	76	0.019	490	0.062	564	0.040	491	0.031	299	0.023

7.10. Study 2 Discussion

The data from Study 2 provide some information on statistical power, non-significant result reporting and publication bias in animal cognition. First, the manually extracted p-value distribution differed from a uniform distribution for two reasons: the cumulative frequency was greater in the observed distribution

for smaller *p*-values ($p < .3$), and also greater for large p-values ($p > .95$). The larger density of smaller *p*-values is consistent with lower powered research in which the null hypothesis was incorrect, but the results did not reach statistical significance. And the density of very large p-values is consistent with researchers applying corrections that might increase *p*-values, such as Bonferroni corrections.

The same pattern was also observed in the StatCheck data, which can be somewhat explained the same two processes. However, any discussion of the StatCheck data must be taken in context with how the data were generated. StatCheck automatically extracts test statistics, degrees of freedom and *p*-values in results written in the text of articles in APA format. *P*-values in tables and figures are not extracted. Therefore, if researchers are more likely to report lower *p*-values within the main text body than higher *p*-values, this could also account for the larger density of *p*-values in the first half of the distribution. Hence, without further investigation, the StatCheck data cannot speak to the relative contribution of reporting biases vs false negative results in determining the shape of the observed *p*-value distribution.

An interesting contrast between the manually and automatically extracted *p*-value distributions is that, unlike in the manual distribution, *p*-values in the range .05 to .10 were much more common than *p*-values in the range .10 to .15 in the Statcheck distribution. This is likely because we extracted results that researchers had interpreted as negative for the manual dataset, but p-values in the range .05-0.1 are often interpreted as “trends” or “marginally significant”. In contrast, the Statcheck dataset automatically extracted all p-values in this range, i.e., irrespective of the authors interpretation of them.

Finally, the Statcheck distribution provides some further evidence of publication, or at least reporting, biases in animal cognition. Figure 2727 contains the characteristic drop in *p*-values just above the significance threshold that is incompatible with any simulated form of research without publication or reporting bias. However, again two sources likely contribute to this drop: genuine publication bias in which p-values above .05 are less likely to be reported in papers than p-values about .05, and less severe reporting biases, in which p-values above .05 are less likely to be explicitly reported in the text body of an article than p-values below .05. Non-significant results reported in tables, or not in APA format would not have been extracted by Statcheck. For example, phrases such as “There was a significant effect of X ($t_{3,6} = 2.56$, $p < .05$), but no other variable, would only have the had the significant effect extracted. Hence, the Statcheck data are unable to provide strong evidence of the specific cause of the drop in *p*-values above 0.05.

7.11. Summary

This Chapter explored negative result reporting and interpretation in animal cognition. In line with previous studies in other disciplines (Aczel et al., 2018; Fidler et al., 2006), non-significant results were often reported as if there were no differences observed in the sample, and this was the case in the titles, abstracts and result sections of papers, although it was most frequent in the titles and abstracts. Because of the distance between statistical hypotheses and theoretical claims, and uncertainty around how no difference statements are interpreted, the consequences of this putative error are uncertain. Nevertheless, these results suggest that researchers should pay close attention to the evidence used to support claims of absence of effects in the animal cognition literature, and prospectively seek to, i) report non-significant results clearly, and ii) use more formal methods of assessing the evidence against theoretical predictions. The *p*-value distribution analysis furthered concerns raised earlier in the thesis that false negative results could be prevalent in animal cognition research and provided some evidence of publication and reporting biases against non-significant results, also. Chapter 8 provides a general discussion of Chapters 5, 6 and 7, and specially comments on the importance but difficulty of secondary data analysis and evidence synthesis projects in animal cognition research.

8. Chapter 8: Barriers to effective evidence synthesis in animal cognition research

Evidence synthesis – the qualitative or quantitative summary and analysis of previous research – is important for cumulative scientific progress (Kousta, 2021), and in this short chapter, I reflect on some of the difficulties in performing effective evidence synthesis of animal cognition research, collating the difficulties I encountered across the synthesis projects of Chapters 5, 6 and 7. I do so by contrasting evidence synthesis in animal cognition with evidence synthesis in medical research. First, I present how evidence synthesis occurs in medical research – a field that often has both the quantity and homogeneity of evidence to perform relatively high-quality systematic reviews and meta-analyses. I next outline six barriers that prevent such effective evidence synthesis across many topics in animal cognition. I conclude by stating that the presence of these barriers should, i) motivate more high-quality and cautious systematic reviews and meta-analyses in animal cognition research, and ii) limit the certainty with which animal cognition researchers present general findings from their research programmes.

8.1. Evidence synthesis in medical research

Systematic reviews and meta-analyses of high-quality randomised controlled trials (RCTs) sit atop of the hierarchy of evidence in medical research (Guyatt et al., 1995). A systematic review aims to synthesise scientific evidence relevant to a certain pre-specified objective. To do so, the review will select studies based on specific eligibility criteria and searches of specific information sources, which determines the scope of the review (Page et al., 2021). Data relevant to the review's objectives are then extracted from these studies and subsequently analysed, either descriptively, or via meta-analysis. To probe the robustness of these analyses, sub-group analyses, meta-regression and heterogeneity tests may be used. Of particular importance is screening studies for risk-of-bias, from which low quality or high risk-of-bias primary studies can be excluded from the summary statistics and meta-analytic estimates (Farrah et al., 2019). While researcher degrees of freedom at each stage of systematic reviews and meta-analyses can limit their authority (Stegenga, 2011), high-quality, registered and transparent systematic reviews and meta-analysis are often recognised as crucial pieces of evidence in scientific research (Cook et al., 1997).

In medical research, high-quality evidence synthesis is facilitated by large amounts of targeted infrastructure and training materials. For example, the PRIMSA guidelines provide a workflow for conducting and reporting systematic reviews (Page et al., 2021), and various organisations offer checklists for risk-of-bias assessments of RCTs, such as the Cochrane Collaboration (Sterne et al., 2019) and the

National Institute for Health and Care Excellence (NICE, 2012). Such checklists can be so useful in medical research because of the similarity of many double-blind, placebo controlled RCTs, independent of the specific drug or illness being studied. For example, biases such as ineffective randomization or non-random attrition can be assessed across every RCT by a skilled coder.

Nevertheless, even with such infrastructure, medical systematic reviews and meta-analyses are often criticised. For example, Stegenga (2011) questioned meta-analyses *de facto* position as the platinum standard of evidence in medicine because of the subjectivity of decisions at each stage of a meta-analysis. And if low quality studies cannot be effectively identified and excluded from meta-analytic estimates, the bias from these studies will also manifest in the meta-analysis. For example, Lawrence et al., (2021) highlighted how flawed trials of ivermectin for the treatment or prophylaxis of COVID-19 were given over 10% of the weighting in two meta-analyses assessing ivermectin's efficacy.

8.2. Evidence synthesis in animal cognition research

In contrast to medical research, systematic reviews and meta-analyses are not yet normal practice in animal cognition research. Now, I highlight six barriers to effective evidence synthesis in animal cognition that researchers should consider when writing reviews.

8.3. Six barriers to effective synthesis in animal cognition

8.3.1. Identifying relevant studies

In medical reviews, systematic reviews of treatments can often identify most relevant studies through key-word searches, often using different iterations of disease/treatment names and study designs (e.g., searching for "randomised controlled trail", "clinical trial", "Phase III trial", etc.). In contrast, systematic reviews in animal cognition face difficulty in defining, i) the topic area being studied, and ii) identifying studies belonging to this topic area. In Chapter 6, for example, I had to develop a key-word search of "social cognition" from key words that experts in the field associated with the topic. Simply searching for "social cognition" and "animal" would have had extremely low specificity and sensitivity (*i.e.,* missed many relevant articles and included many irrelevant articles). This could have been remedied by narrowing the scope of the review, e.g., to a particular paradigm. However, such a process, i) requires the paradigm to have been consistently named in the past, and ii) would miss many studies relevant to the cognitive ability in question that did not use the specific paradigm.

Another alternative would be to conduct hand searches of journals or contact experienced senior colleagues, but this still requires subjective expert judgements on which studies to include and/or access

to these experts. Moreover, as much animal cognition research is published in non-specialist journals without designated animal cognition sections, hand searches may be too resource intensive to be viable.

8.3.2. Study design heterogeneity

Heterogeneity is a major barrier of quantitative evidence synthesis in animal cognition. As outlined in Chapter 3, many studies in animal cognition sample single instances of treatments, measurements and settings, and a small number of experimental units. When synthesising data from across such these studies, researchers should consider whether the studies they have identified are homogeneous enough for average summary statistics, or meta-analytic estimates, to be meaningful. This should be judged on a case-by-case basis, and can be facilitated by performing sub-group analyses. Notably, this critique also seems to ring true for small-scale RCTs, that might sample a single treatment in a single setting (RCTs often measure several outcome variables, however). However, in the case of RCTs, we may often have a greater mechanistic understanding of how a drug or an indication works (e.g., Jones et al. 2021), and hence which methodological differences may be effect-modifying.

8.3.3. A small number of studies

The problems with study heterogeneity are exacerbated with a small number of studies, because the degree of heterogeneity becomes difficult or impossible to detect. Moreover, just as statistical analyses of individual experiments, too small a sample of studies will preclude informative meta-analysis or phylogenetic models, i.e., they will produce wide or indefinable confidence intervals. Importantly, the number of studies on any given topic in animal cognition *within* an individual species or group will normally be low, with exceptions for particularly popular tasks or species.

8.3.4. Reporting heterogeneity and low data availability

Even when there are enough studies on a certain topic, differences in how the methods and results of these studies are reported can be a barrier to effective evidence synthesis. In Chapter 6, my coders – experts in animal cognition research - had difficulty identifying whether experimenters or second observers were blinded during the studies, and even difficulty identifying how the behaviours were measured. Thus, even if the raw number of studies is large in a systematic review project, the amount of useable data may be lower for many outcomes. If this attrition is non-random enough, this will also bias the results of the review project. Moreover, reporting heterogeneity can prevent effective quantitative synthesis of results. For example, if effect sizes are not reported alongside statistical tests, and these cannot be computed from the summary statistics and no raw data are available, this can reduce the data

available for meta-analysis. This type of issue was seen across Chapters 5, 6 and 7, in which papers that did not report exact *p*-values reduced the data available for the *p*-value distribution analyses.

8.3.5. Publication bias and lack of study registration

Insofar as the aim of synthesis projects is to assess all available evidence on a certain topic, publication bias is a major issue for synthesis projects in animal cognition. As the extent of publication bias is unknown for most topic areas in animal cognition, attempting to measure publication bias should be a feature of all systematic reviews in animal cognition research. However, there is no fool-proof method to detect this. In this thesis I used two methods, assessing the nature of the claims published in the literature and the *p*-value distributions of hypothesis tests, and part of Chapter 9 presents a final measurement attempt through self-report surveys. These could be complemented by statistical tests such as Egger's regression test on funnel plots, but these tests suffer from low power at small sample sizes and can give significant results even in the absence of publication bias (Sterne et al., 2011).

8.3.6. Difficulty of assessing individual study risk of bias

Perhaps the largest barrier to effective evidence synthesis in animal cognition is the relative difficulty of identifying studies at high risk of bias. Because of the relative homogeneity of double-blind placebo-controlled RCT research, known biasing factors are well understood and often well reported, with checklists designed for this purpose (Sterne et al., 2019; NICE 2012). In contrast, checklists have limited utility in assessing the quality of animal cognition studies – because, in addition to issues of randomisation and blinding, tests of animal cognition are usually poorly validated (see Chapter 4). To assess the quality of any individual study would therefore require an in-depth post-publication peer review of the study, and even with this there is no guarantee that high-risk-of bias (or confound) studies will be identified. Such projects, although labour intensive, are possible: for example in cognitive psychology, Vater et al. (2021) performed a critical systematic review of the Neurotracker perceptual-cognitive training tool in which all 16 studies fitting their minimum quality criteria had some level of review performed.

8.4. Examples of evidence synthesis in animal cognition

The six barriers presented above aside, systematic review and meta-analysis projects have strong potential to improve evidence assessment, communication, and scientific thinking in animal cognition research. I now give examples on how some of these barriers have been overcome in pieces of evidence synthesis.

8.4.1. Reviews based on specific paradigms

Reviews of specific paradigms go someway to mitigating the effects on heterogeneity on meta-analyses. For example, Clark et al. (2019) performed a meta-analysis focusing on the object choice task, a task in which researchers test animals' ability to respond to directional cues, such as eye gaze or pointing, to an object. They used a keyword search of the literature on object-choice task performance, and, in response to a lack of data on some species, focused their analysis on canines and primates. In addition to the keyword search, they searched the references of published articles and their own knowledge to include other, relevant studies – highlighting the difficulty of conducting thorough keyword searches in animal cognition. Through the review and meta-analysis, they were able to show how task designs systematically varied between the two groups – notably regarding the presence (primates) or absence (canines) of a barrier, and a greater inter-object distance for canines. These data raise the possibility that the previously claimed superiority of dogs at the object choice task might be due to methodological, rather than taxonomic, differences (see also Clark & Leavens, 2019). For a similar example of a meta-analysis of a specific paradigm, see Qu and Kwok's (2020) meta-analysis of animal uncertainty monitoring, focusing on the "opt-out" task.

8.4.2. Reviews based on general questions

Reviews of larger areas of animal cognition research will likely encounter large amounts of difficult-to-explain heterogeneity. For example, in the meta-analysis on the link between personality measures and learning ability, Dougherty and Guillette (2018) found a high level of between-study heterogeneity, which several proposed moderator variables could not effectively account for. From their meta-analysis, Dougherty and Guillette reported an overall mean effect size that was not significantly different from zero ($r = 0.098$), but the modulus of the mean effect size $|r|$ was 0.268, and significantly different from zero. These data seem consistent with Dougherty and Guillette's claim of a "small but significant" relationship between variation in personality and variation in learning across species in the absolute scale, but with a variable direction. However, whether this can be teased apart from sampling error across the individual studies is an open question and highlights the interpretation difficulties that high heterogeneity brings.

8.5. Summary: More systematic reviews and meta-analyses are necessary to understand evidential quality in animal cognition research

Overall, systematic evidence synthesis is a difficult process in animal cognition. Researchers must balance the need to isolate homogenous enough bodies of data so that they are meaningful when aggregated, while keeping these data relative to the original aim of the review. Nevertheless, animal

cognition research stands to gain a lot from systematic review and secondary data analysis projects. While the conclusions of these studies will contain many caveats, as were necessary in Chapters 5, 6 and 7, these studies can help to: i) focus critique at the level of research programmes; ii) help assess the risk-of-bias; and iii) understand and communicate the limits and uncertainty of the data that animal cognition research produces.

9. Chapter 9: Attitudes toward bias, replicability and scientific practice: A survey study[13]

Throughout this thesis I have raised concerns about replication, bias and incentives in animal cognition research, as well as our ability to detect and study this. While these issues are sometimes discussed in animal cognition, these debates are often performed by a minority of stakeholders in animal cognition research—often between those who claim discoveries of "higher" processes in animals and their corresponding 'killjoys' or skeptics, accompanied by a meta-commentary from a small number of interested researchers and philosophers (for example Allen, 2014; Anderson & Gallup, 2015; Barrett, 2015; Craig & Abramson, 2018; Despret et al., 2016; Eaton et al., 2018; Farrar & Ostojic, 2019; Heyes, 2015; Leavens et al., 2019; Penn & Povinelli, 2013; Povinelli, 2020). But how effectively these debates are reaching animal cognition researchers in general, and how they are received, has garnered little attention.

Survey studies can address this by directly asking researchers their opinions on key debates in the field, how their own research practices are shaped by these debates, and what they feel is incentivised in academia. For example, survey studies have quantified the negative effects on researchers' mental health due to academia's "publish or perish" culture (Haven et al., 2019), and researchers often report that scientific incentives are misaligned with their scientific ideals. For example, ecology researchers reported that while they thought replication studies were a crucial use of resources, they experienced difficulty obtaining funding for them and, even if they were performed, they perceived barriers to publishing them (H. Fraser et al., 2020). More directly, researchers have self-reported using false-positive inflating research practices at non-negligible rates (Agnoli et al., 2017; Fiedler & Schwarz, 2016; H. Fraser et al., 2018; John et al., 2012), and also measured editor and reviewer biases against replication studies (Neuliep & Crandall, 1990, 1993).

In the current study, I invited 1001 researchers who have published in animal cognition journals in the last three years to answer a range of questions about bias and research practices in animal cognition research. The survey consisted of five blocks of questions, broadly covering, i) bias, ii) publication practices, iii) statistics, iv) replication, and v) how researchers derive their own beliefs about animal

[13] This chapter contains material published in Farrar B.G., Ostojić L., Clayton N.S. (2021) The hidden side of animal cognition research: Scientists' attitudes toward bias, replicability and scientific practice. PLoS ONE 16(8): e0256607

cognition. These topics were based around the issues raised in this thesis, and informed by wider debates in animal cognition and scientific reform more generally.

The survey had three aims, namely, i) to survey the extent to which researchers are concerned about certain research and publication practices in the field, ii) to collect direct evidence of the rates of these practices from researchers themselves, practices that may otherwise be difficult to observe, and iii) to provide researchers with the opportunity to voice any concerns or opinions they have about how animal cognition research operates. These data may impact the field in three ways. Firstly, they can help researchers critically evaluate the evidential strength of published findings, given how frequently researchers estimate certain biases to be present. Secondly, they can facilitate debates on the effectiveness of the scientific process in animal cognition and engage researchers and students in these debates. Finally, they can help to identify barriers to effective scientific research of animal cognition that can inform policy making in journals, funding bodies and hiring committees, as well as decision making by individual animal cognition researchers.

9.1. Methods

9.1.1.Sample

I invited all researchers who are a first, last or corresponding author on any type of article published in the past three years (i.e., 2018-2020 inclusive) from the following six animal cognition journals to complete our survey: *Animal Cognition*, *Animal Behavior and Cognition*, *Journal of Comparative Psychology, International Journal of Comparative Psychology, Journal of Experimental Psychology: Animal Learning and Cognition, Frontiers in Psychology: Comparative Psychology.* I viewed every article from these journals between 2018 and 2020, and extracted the email addresses of the first, last and corresponding authors. If these email addresses were not provided in the article, BGF conducted a keyword-based web search to try to find one for the author in question. In total, 1161 authors were identified and email addresses for 1004 of these could be located from the articles or web searches. Of these, three email addresses were those of my research team on this project, leaving a final sample of 1001. Emails were sent to these 1001 researchers in January 2021. Sixty-four emails returned errors, and BGF conducted further web searchers to identity alternative emails for these researchers, of which 32 were obtained and the survey invite emailed to. Of the 969 successfully sent emails, 210 completed surveys were returned (response rate = 21.6%).

Researchers completed a questionnaire hosted on Qualtrics. The study protocol was approved by the University of Cambridge's Psychology Research Ethics Committee (PRE.2020.096). The survey was piloted on several volunteers from the Comparative Cognition Laboratory at the University of Cambridge. The full survey is detailed below, and the anonymized survey data and analysis code are available at osf.io/6j7kp.

9.1.2.Survey

The exact survey is presented in Figures 28 to 32. The five blocks covered topics as follows: The **Bias** block asked researchers about experimenter bias and objectivity in their own work, and about the role bias might play in shaping the results and theories in animal cognition research more broadly. The final topic of the bias block was Morgan's canon—the notion that animal behaviour should not be interpreted in terms of "higher" psychological processes if it can be fairly interpreted in terms of "lower" processes—with researchers answering whether they agreed that "Morgan's canon is important to use when interpreting the results of animal cognition research". The **Publication** block first asked to what extent researchers thought that they themselves, and other researchers, make appropriate claims when submitting research for publication. Second, as a direct measure of publication bias, I asked researchers which proportion of their own studies has been published, or will be published for ongoing studies, as well as the reasons why some of their studies go unpublished. The **Statistics** block then measured researchers' confidence in their own statistical analyses, and their ability to judge the validity of other analyses. Next, it asked researchers to estimate the prevalence of "questionable research practices", which may increase the likelihood of spurious findings in their own and in others' research. The **Replication** block first focused on attitudes towards replication studies; how important are replications, and are replications performed often enough in their own area of research and others? Second, it asked researchers about whether they believe their own area of research, or other areas of research in animal cognition, would experience a 'replication crisis' if multiple replication studies were attempted, and how many of these replication studies they would predict to be 'successful'. Finally, the **Belief** block asked researchers a range of questions about how they decide what to believe about animals' cognition. I asked researchers about the role that scientific experiments and day-to-day experience play in shaping these beliefs, as well as how often they agree with the conclusions presented in scientific papers.

This section involves questions about **publications** in animal cognition.

When publishing a paper, researchers make claims about what their data mean.

When submitting a paper for peer review, I tend to:

- ○ make weaker claims than are warranted by the data
- ○ make appropriate claims given my data
- ○ make stronger claims than are warranted by the data
- ○ N/A

After peer review, my claims tend to:

- ○ become weaker
- ○ stay the same
- ○ become stronger
- ○ N/A

When submitting papers, I believe that **other researchers** tend to:

- ○ make weaker claims than are warranted by the data
- ○ make appropriate claims given their data
- ○ make stronger claims than are warranted by the data

Publication bias occurs when the decision to publish studies is dependent on the results of the studies, e.g., when negative results are not published.

What **percent** of the studies that you have performed have been published and/or you think will be published?

0 10 20 30 40 50 60 70 80 90 100

% Of studies

Figure 28: The Publications Questions

Figure 28 (cont'd): The Publications Questions

In the past, have you been unable to publish a study you have performed, or have you ever decided not to publish a study?

- ○ Yes
- ○ No
- ○ N/A

Why did you choose not to publish the study, or why do you think the study was not published?

Do you have any other comments about **publication** in animal cognition?

This section asks about **bias** in animal cognition.

	Never	Rarely	Sometimes	Often	Always	N/A
When performing research, I find myself hoping for one result over others.	○	○	○	○	○	○
I am concerned that I might bias the results of my studies towards certain results.	○	○	○	○	○	○
I can detach from any biases to perform objectively fair tests of animal cognition	○	○	○	○	○	○

	Strongly Disagree	Disagree	Neither agree nor disagree	Agree	Strongly Agree	N/A
The results and theories **in my area** of animal cognition are strongly affected by researchers' biases	○	○	○	○	○	○
The results and theories **in other areas** of animal cognition are strongly affected by researchers' biases	○	○	○	○	○	○
If I knew the topic and the authors, I would be able to guess the conclusions of a published study without reading it	○	○	○	○	○	○

Figure 29: The Bias Questions

Figure 2 (cont'd): The Bias Questions

Morgan's canon states that: "In no case is an animal activity to be interpreted in terms of higher psychological processes if it can be fairly interpreted in terms of processes which stand lower in the scale of psychological evolution and development."

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree
Morgan's canon is important to use when interpreting the results of animal cognition research	○	○	○	○	○

Do you have any other comments about **bias** in animal cognition research?

This section involves questions about **replications** in animal cognition research.

Consider 100 typical papers in your research area. What percentage of studies would successfully replicate the original results, if the same protocol was used on a new sample of similar size in the same species (assuming that this was possible)?

Use whichever definition of successful you think is appropriate.

0 10 20 30 40 50 60 70 80 90 100

% Successful

Please indicate the percentage of studies you think would successfully replicate the original results if replication studies had a **sample size of 1000 animals**?

0 10 20 30 40 50 60 70 80 90 100

% Successful

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree	N/A
My area of animal cognition research would experience a "replication crisis" if attempts to replicate most of its studies were conducted.	○	○	○	○	○	○
Some areas of animal cognition research would experience a "replication crisis" if attempts to replicate most of its studies were conducted.	○	○	○	○	○	○
I could identify animal cognition studies that would successfully replicate and those which would not.	○	○	○	○	○	○

Figure 30: The Replicability Questions

Figure 30 (cont'd): The Replicability Questions

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree	N/A
It is important that replication studies are performed in animal cognition research.	○	○	○	○	○	○
Enough replication studies are published **in my area** of animal cognition.	○	○	○	○	○	○
Enough replication studies are performed in animal cognition in general.	○	○	○	○	○	○

Do you have any other comments about **replication** in animal cognition research?

This section involves questions about **statistics** in animal cognition research.

	Strongly disagree	Somewhat disagree	Neither agree nor disagree	Somewhat agree	Strongly agree	N/A
When I perform a statistical analysis, I know that the analysis is valid and appropriate.	○	○	○	○	○	○
I could explain why my analysis is appropriate and valid to another researcher.	○	○	○	○	○	○

	Never	Rarely	Sometimes	Often	Always
When I read or review an article in my area of animal cognition, I can assess whether the statistical methods used are valid and appropriate.	○	○	○	○	○

Figure 31: The Statistics Questions

Figure 31 (cont'd): The Statistics Questions

There is growing concern that researchers use false positive inflating research practices in science, such as:

- Performing many analyses and selectively reporting the statistically significant ones
- Reporting an unexpected finding as if it was predicted from the start
- Data dredging/p-hacking/fishing for significance
- Selectively excluding data points to produce a significant/desired result
- Collecting more data until a significant/desired result is obtained

---

How common do you think these research practices are in...

	Never	Rare	Sometimes	Often	Always	N/A
Your own research	○	○	○	○	○	○
Other research in animal cognition	○	○	○	○	○	○

Do you have any other comments about **statistics** in animal cognition?

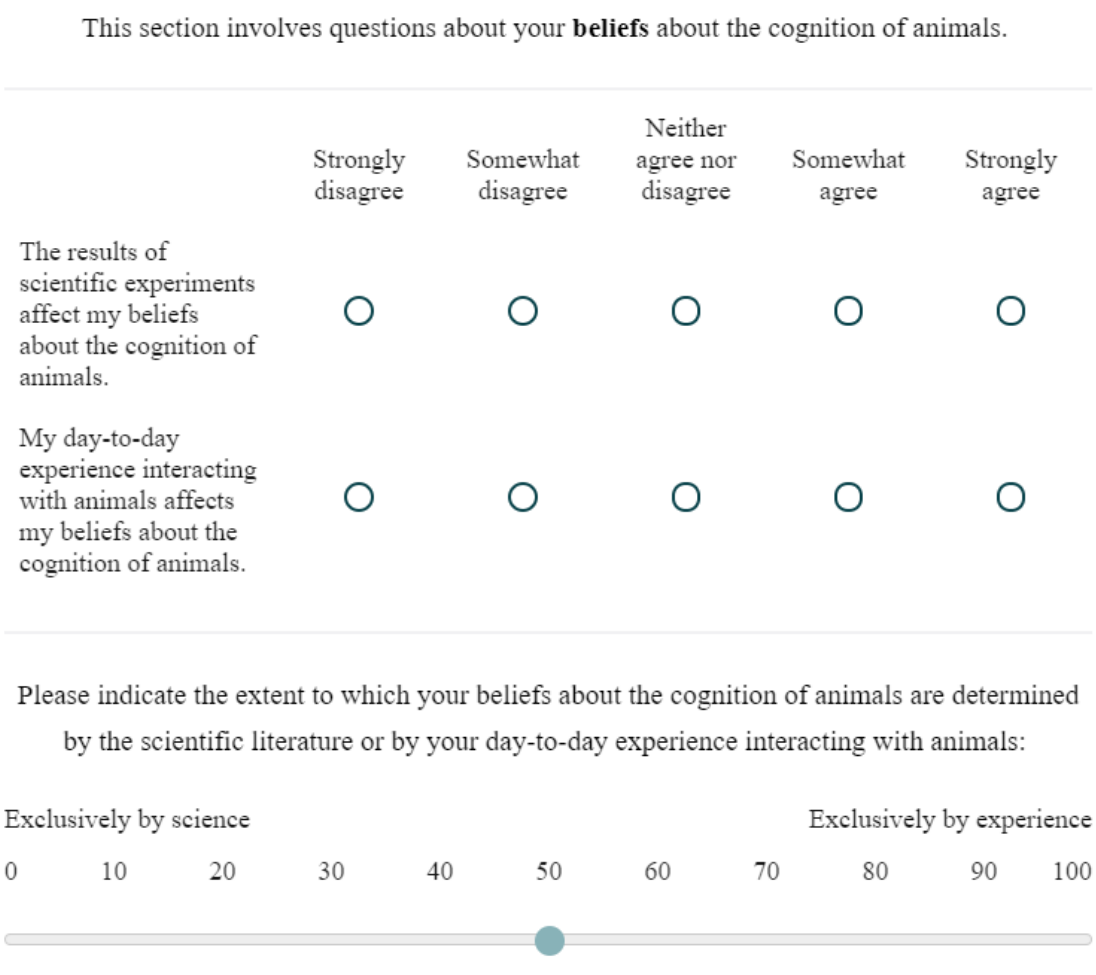


Figure 32: The Belief Questions

Figure 32 (cont'd): The Belief Questions

	Never	Rarely	Sometimes	Often	Always	N/A
When I read a paper in **my area** of animal cognition research, I agree with the authors' interpretation of their data	○	○	○	○	○	○
When I read a paper in **other areas** of animal cognition research, I agree with the authors' interpretation of their data	○	○	○	○	○	○

Do you have any other comments about your beliefs about the cognition of animals and the role of science in forming them?

9.1.3.Analysis

Throughout the results, I provide direct quotes of participants' answers to the free-text responses. These quotes were taken from participants who, at the end of the survey, opted in for their free-text answers to be shared openly and were screened for any identifying information. If a free-text response contained clearly identifying information, it was excluded from the open dataset. All the free-text answers for which I received consent to share, and which did not contain identifying information are openly available at osf.io/6j7kp. In addition to directly quoting participants' free-text answers, of which only a minority could be included in the report, I also categorized their free-text responses based on the common themes that they included within each block. First, I read through all responses and identified common themes in participants' responses. I then marked whether each response fit each category or not. If a response matched more than one category, this was still recorded, i.e., a single response could in principle fit all the categories. A second author (LO) was given the category descriptions and, blind to the first coder's decisions, also marked whether each response fit each category or not. Of BGF's 481 decisions to label a response with a category, LO independently agreed with 402 (83.6%) of them. In addition, LO made 103 classifications that BGF had not originally and suggested four further category labels, three of which were retained. Each disagreement was resolved by discussion between LO and I, with the most disagreements either being an error from one of the two coders originally, or cases where both coders agreed that the statement was ambiguous, i.e., there were no cases of disagreement that could not be resolved through discussion. The category-based analyses are presented for the Publication, Statistics, Replication and Belief blocks. For the Bias block, I chose to split the results of the open-ended question ("Do you have any other comments about bias in animal cognition research?") into two tables, as participants' free-text responses were split between providing examples of biases in animal cognition research and elaborating on their Likert-type responses to the question about Morgan's canon. In addition to the category based-analysis, I also present some quotes in-text that I feel highlighted an important topic that our category-based analysis might have missed. Where some themes occurred across blocks of topics but were not necessarily directly related to the topic in question, I present these in a "miscellaneous" section, although this was not performed systematically.

9.2. Results

9.2.1.Demographics

From 1001 invitations, I received 210 completed surveys (response rate = 21.6%). The sample of researchers had published a median of 17 papers on topics in animal cognition (IQR: 8 – 50) and had been active in the field for a median of 14 years (IQR: 8 – 25). Table 21 displays these demographics.

Table 21: The number of papers published and years active in animal cognition of the 210 researchers completing the survey.

Number of papers	**0**	**1-5**	**6-10**	**11-25**	**25-75**	**> 75**
%	0.4	17.6	22.8	19.5	24.8	14.8
N	1	37	48	41	52	31
Years active	**0**	**1-3**	**3-7**	**8-15**	**15-25**	**>25**
%	0.4	2.3	18.1	28.1	21.4	20.5
N	1	5	38	59	45	43

One response for years active was left blank and therefore excluded. The one researcher who reported 0 for papers published and years active later described publishing in at least one of our target articles, suggesting that the 0 responses may have been in error.

9.2.2.Bias

I asked researchers about bias in their own experiments, and their perceptions of bias across the field. Researchers frequently reported either sometimes (39.7% of respondents) or often (38.8%) hoping for one result over another when performing research, and researchers were split between either being rarely concerned (36.5%) or sometimes concerned (30.3%) that they might bias the results of their studies towards a certain conclusion. Nevertheless, they reported that they could often (45.8%) or always (38.4%) detach from any biases to perform objectively fair tests of animal cognition (Figure 33).

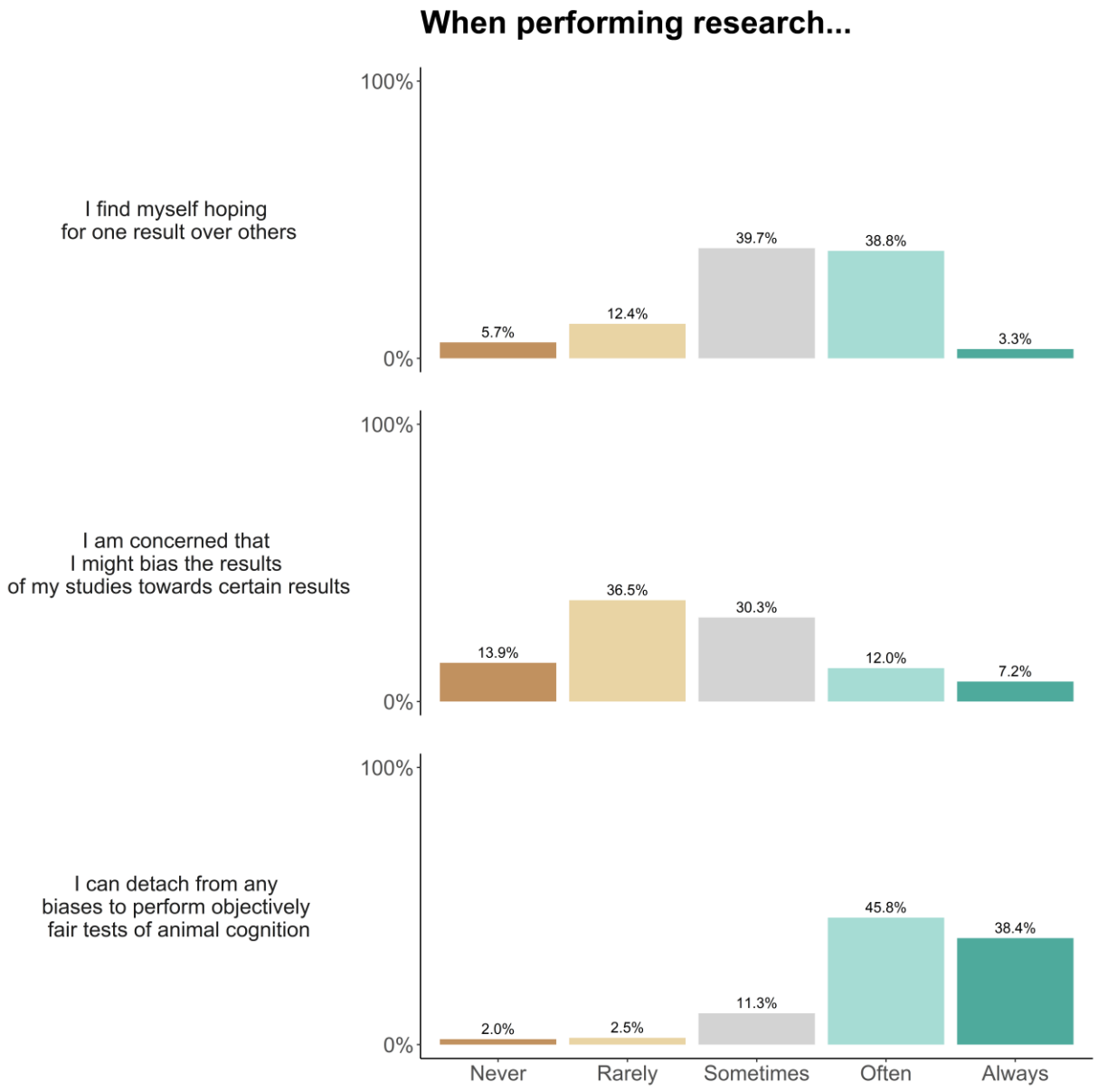


Figure 33: Animal cognition researchers' self-reported concern about bias in their own studies (N = 210). Percentages may not add to 100% due to a small number of NA responses.

In terms of bias across the field, researchers were split between agreeing (29.6%), disagreeing (23.8%) and neither agreeing nor disagreeing (36.4%) that the results and theories in their own area of animal cognition are strongly affected by researchers' biases. Responses were similar when researchers were asked to consider bias in other areas of animal cognition, but more researchers agreed that the results and theories are strongly affected by researchers' biases (agree: 36.0%; neither agree nor disagree: 39.0%; disagree 14.5%). Researchers were split between agreeing (34.0%), disagreeing (22.3%) or neither

agreeing or disagreeing (30.6%) that if they knew the topic and the authors, they would be able to guess the conclusions of a study without reading it (Figure 34). Notably, most respondents tended to avoid the extreme responses—no more than 10.5% of respondents chose the strongly agree or strongly disagree for these questions on bias.

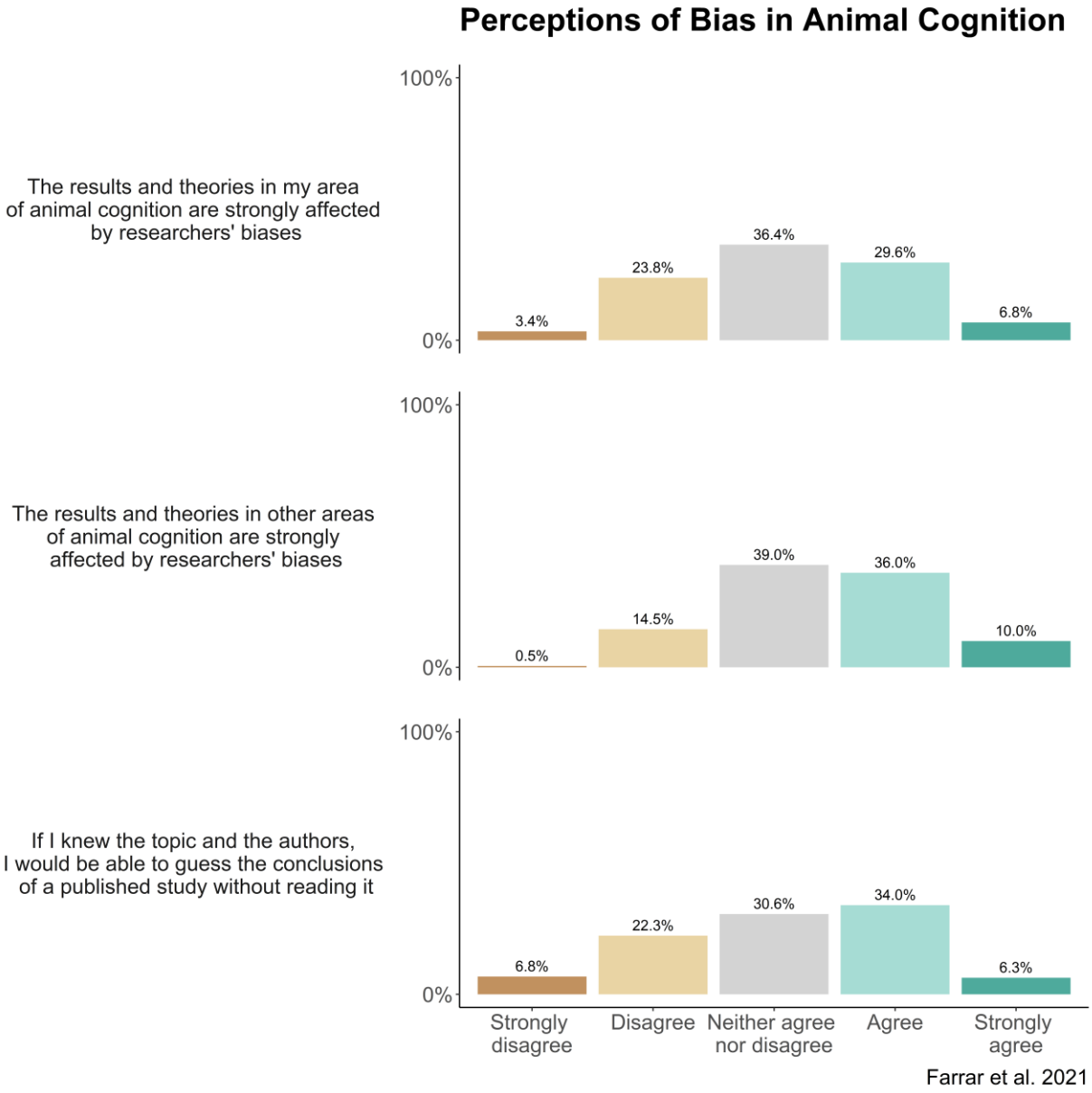


Figure 34: Animal cognition researchers' self-reported concern about bias in animal cognition research (*N* = 210).

I received 68 free-text responses concerning bias in the field, many of which elaborated on the question about Morgan's canon. However, researchers reported a diverse range of attitudes towards bias

in the field. While most researchers reported they could detach from their own biases readily on the Likert-measure, perhaps through using measures such as blinding, other researchers expressed skepticism about the ability to perform research objectively:

"As to the first three questions on my own bias - it is NEVER possible to detach yourself from your own biases. You can only try your best and take as many steps as possible to control for this, which I do… As to hoping for one result over another - as negative results are unpublishable, any sane scientist will hope for positive results. Our careers, and often our livelihoods, rely on getting positive results and publishing them. Too much is at stake to pretend that there is no bias."

Researchers indicated several different forms of bias that might affect animal cognition research, ranging from anthropomorphism and confirming "higher" abilities in animals, to excessive skepticism. Table 22 presents a selection of these reported biases.

Table 22: Animal cognition researchers' beliefs about bias in animal cognition research

Do you have any other comments about bias in animal cognition research?	N	Exemplars
Provided an example of bias	31	*"Bias is of two kinds: (a) bias against animals in comparison with humans (pro-human bias) and (b) bias to interpret animal behaviour as evidence for complex cognition (pro-animal bias). Both kinds of bias undermine the legitimacy of animal cognition research."*
Suggested that bias is predominantly in how the data are interpreted	30	*"The bias I typically see is a bias to produce a narrative that goes beyond the data, I am not sure whether that bias goes into the design/approach that produces the data. I almost never have concerns that design/data might be unethically tweaked (i.e., conscious bias)."*
Suggested that bias is inherent to many study designs	29	*"It is often in the selection of the behavioral markers that the biases are most strongly evident, so the biases are well-entrenched well before data analysis. Given the modern trend to only provide heuristic descriptions of what was measured and reliance on inter-observer reliability to justify those measurements means that the rationale used for the selection process is usually hidden from the reader. This makes it difficult to identify the implicit biases in the study."*

9.2.3.Morgan's canon, simplicity and parsimony

I next asked researchers about the role of Morgan's canon. Most researchers agreed somewhat (38.6%) or strongly (31.9%) that Morgan's canon is important to consider when interpreting the results of animal cognition research (Figure 35). However, researchers often elaborated on these answers in the free text responses, revealing a more nuanced perspective of the use of Morgan's canon, which are detailed in Table 23.

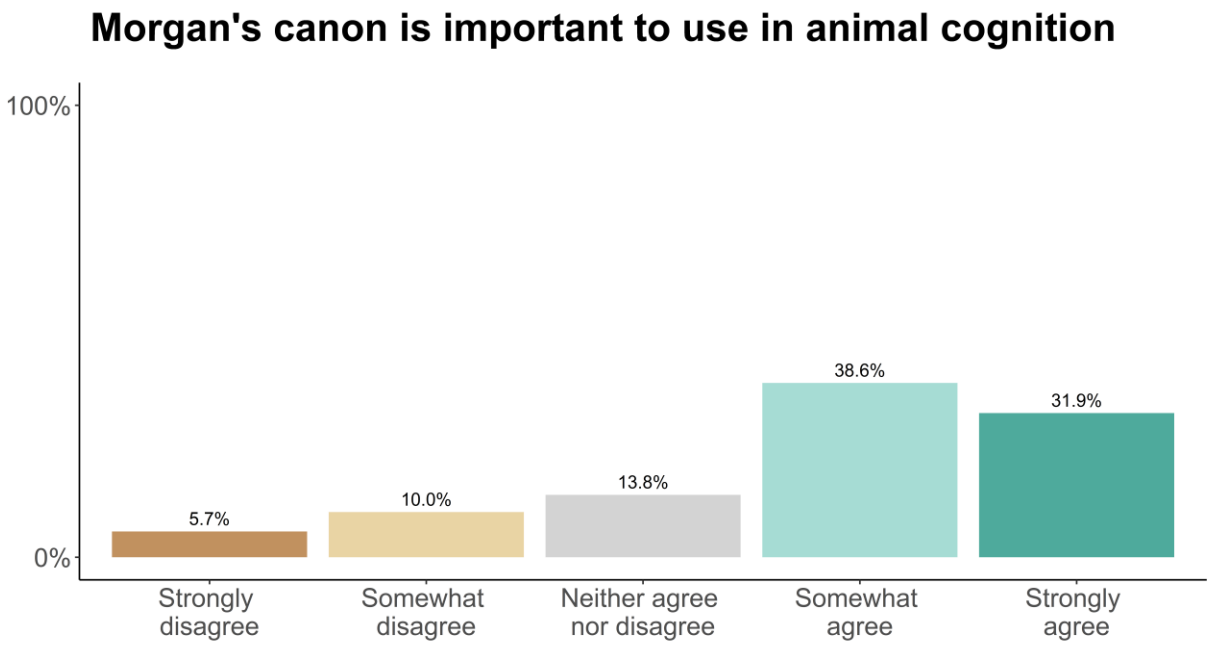


Figure 35: Animal cognition researchers' endorsement of Morgan's canon (N = 210).

Table 23: Animal cognition researchers' attitudes towards Morgan's canon as a tool in animal cognition research.

Do you have any other comments about bias in animal cognition research? Answers that referenced Morgan's canon	N	Exemplars
Caveated or criticized the use of Morgan's Canon	16	*"In my opinion, Morgan's canon often leads to more bias rather than less. As scientists, it is important to have an open mind in both directions. E.g. looking at the evolutionary tree of a species when interpreting its behaviour is often more conclusive that Morgan's canon."* *"I think that biases on the cognitive processes underlying certain behaviours can go both ways. One could overstate the complexity, as much as one could underestimate it. That is why in general I do not consider Morgan's canon to be always useful: to use it best, we would need to have a clear understanding of what process "stand lower in the scale of psychological evolution" without pre-existing biases."*
Suggested that parsimony is important when interpreting data	30	*"In terms of Morgan's canon, it is not dissimilar to parsimony in phylogenetics or Occam's razor in normal scientific inquiry. Showing skepticism in cause does not suggest that more complicated cognitive explanations exist, but the onus is on the researcher to demonstrate."* *"While Morgan's canon is useful as a philosophical tool, I do think that it often conflicts with parsimony arguments made from phylogeny, so in practice I feel it often does not help per se."*

9.2.4.Publication

I asked researchers whether they believe themselves and others to make appropriate claims when submitting research for publication, and how many of their studies end up being published. When submitting papers for publication, 86.0% of researchers reported that they make appropriate claims given their data, while only a small number stated that they overclaim (7.7%) or underclaim (5.8%). In contrast, our sample was split between believing that other researchers were likely to make stronger claims than warranted by their data (56%), and believing that others make appropriate claims (43%, Figure 36). Researchers reported that their own claims usually stayed the same (69.0%) or became weaker (21.0%) after peer review. A minority of researchers reported that their claims increased in strength (9.0%). When asked how many of their studies had, or for ongoing studies, will, end in publication, the median response was 80% (IQR: 70% - 90%, Figure 37Figure 37). However, there was a large spread in responses, with 23 respondents saying 50% or fewer of their studies have been published, and 17 reporting that they have published all of their studies.

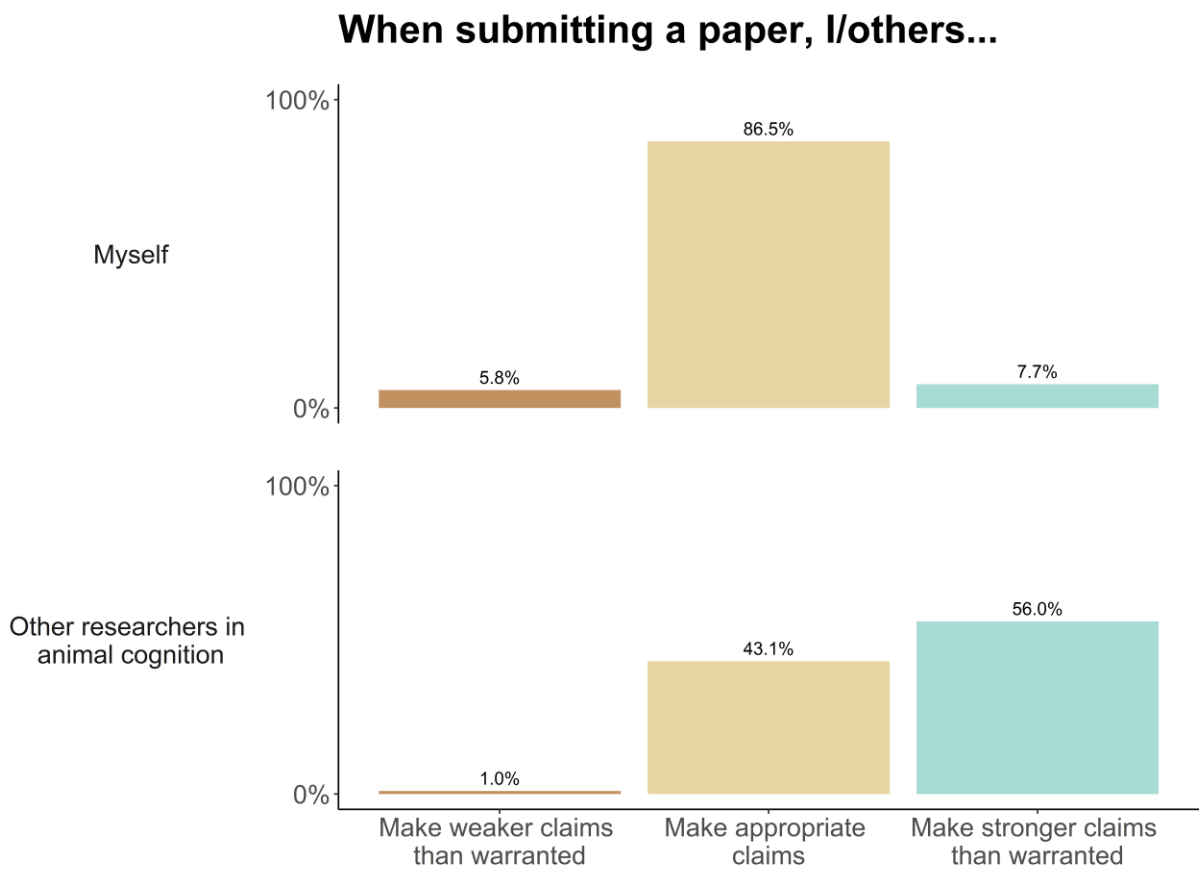


Figure 36: Animal cognition researchers' beliefs about overclaiming and underclaiming when submitting research articles for publication, N = 210.

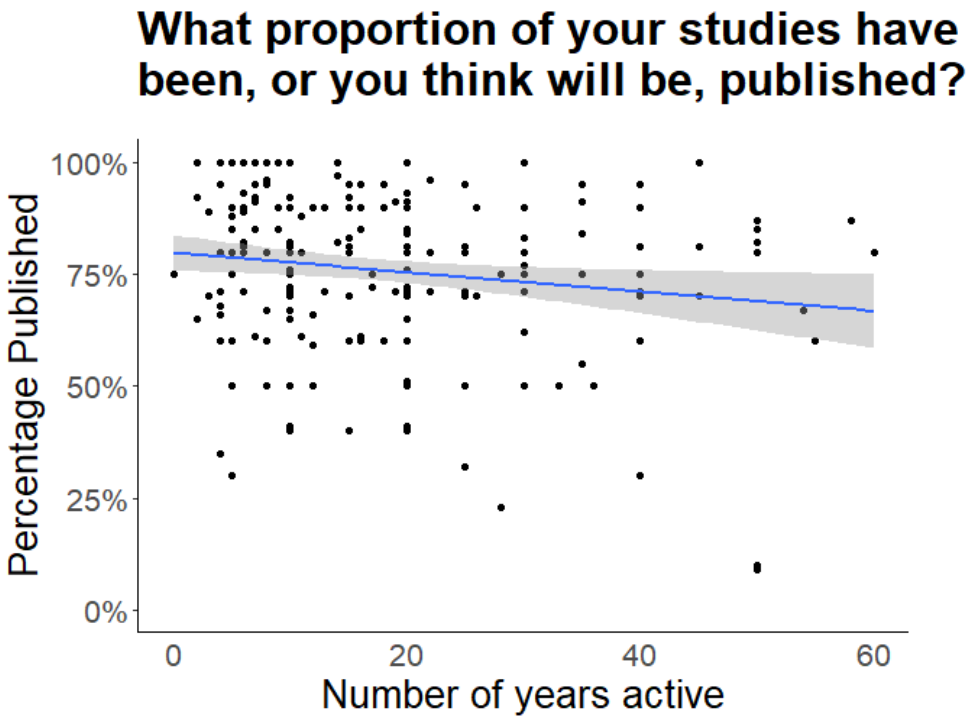


Figure 37*:* Animal cognition researchers' self-reported proportion of studies that they have run and then published, by their-self reported number of years in the field. *N* = 208.

I received 144 free-text responses from researchers explaining why certain studies of theirs had not been published. The responses suggested several different causes of publication bias in the field. Some researchers reported self-filtering studies they deemed of little importance:

> *"I have a few studies that are just not adequate to publish, in terms of experimental design, subject size, or no informative findings (and I'm including null results as potentially informative). These are my own issues, not that of the publication process."*

Another reported cause of self-driven publication bias was a lack of incentives to publish all research, either due to time constraints or perceptions of how publishing all work would affect funding opportunities:

> *"My position is dependent on grant funding, this contingency is coercive to publishing only the studies that strengthen the grant."*

Although not one of our identified themes, sixteen researchers (11% of free-text responses) also reported that publication bias was enforced by journals, reviewers and editors:

“Consistent rejection across journals, which typically reported that the findings were not "attractive enough" (e.g., replications, inconclusive results, etc.)”

Through the categorisation analysis, the most common themes we identified were articles not being published for containing inconclusive results (31), design limitations (30), negative results (29), insufficient resources for publication (29) and too few data (28). In Table 24, I highlight quotes from each of the 10 themes we identified in the responses. Next, Table 25 highlights several quotes from the open-ended free-text question about publication practices in animal cognition.

Table 24: Animal cognition researchers' explanations for why some of their studies go unpublished.

Why was your animal cognition study not published?	N	Exemplars
Inconclusive results	31	*"The results were not so clear to discuss something."*
Design limitation	30	*"Something about experiment was faulty so didn't bother submitting. Or, something was pointed out in peer review that made the study seem not worth trying to publish."*
Negative or uninteresting results	29	*"It was just too difficult to get "negative results" past referees."*
Lack of resources	29	*"Mostly [due to] time to write the studies up. I have too many on my "to do" list"*
Too few data	28	*"Not enough data for reliable conclusions"*
Unreliable data	17	*"Because the design was weak, the experimenter unexperienced… I just wasn't sure whether to trust the data and I did not want to publish any potential false positive/negative findings"*
Reviewer bias	13	*"Theoretical rivals killed the publication because the outcomes didn't fit with their theory"*
Irrelevant data	8	*"The data were incomprehensible, and it appeared the animals failed to learn anything related to the task."*
Training failure	5	*"I often decide not to publish studies if the animals were unable to train to the basic level required to complete the study"*
Replication studies	2	*"Non-significant results or that only previously published outcomes were replicated"*

Table 25: Animal cognition researchers' opinions on publication practices in the field.

Do you have any other comments about publication in animal cognition?	N	Exemplars
Highlighted the difficulty of getting negative results published	15	*"It is next to impossible to publish negative results in animal cognition."*
Highlighted the difficulty of interpreting negative results	7	*"Studies with negative results often needs additional controls to show it is a true negative; most often animal cognition studies are initially designed to control for that a potentially positive result is a true positive. There are many more ways for something to be negative than to be positive, therefor particular care must be given when publishing such data (negative or no results can often be the result of a bad design)."*
Highlighted an excessive focus on publishing "exciting" or "novel" results	7	*"There are still strong incentives towards publishing "wow!" findings showcasing supposedly "clever" or "human-like" abilities."*
Lack of time to publish everything	3	"I just have not had time to publish them."
Other/Other barriers to publishing in animal cognition	15	*"There is a constant pull of the wishful thinking. If we let this go on unchecked, it will eventually converge on what people already think is true and/or what they wish were true."*

9.2.5.Statistics

I asked researchers about their confidence in their own statistical analyses, their ability to assess others' analyses, and the rate of questionable research practices in the field. Researchers strongly or somewhat agreed that when they perform a statistical analysis, they know it is appropriate and valid (strongly agree: 53.2%, somewhat agree 42.9%), and that they could explain why this was the case to another researcher (strongly agree: 59.5%, somewhat agree 36.6%, Figure 38). When reading or reviewing others' research, our sample reported that they could often (59.8%), sometimes (23.4%) or always (12.4%) assess the validity of the analysis. A minority of researchers reported that they could rarely (3.8%), or never (0.5%) assess the validity of the analysis. When asked how often they themselves or other researchers performed questionable research practices (QRPs), which may induce false positive findings,

researchers reported that they themselves rarely (41.1%), never (31.2%), or sometimes (20.3%) conducted QRPs. However, researchers thought that others either sometimes (52.7%), often (27.9%), or rarely (18.4%) did so (Figure 39). I received 66 free-text responses about the use of statistics in the field, from which we identified 13 general themes. These themes are highlighted in Table *26*.

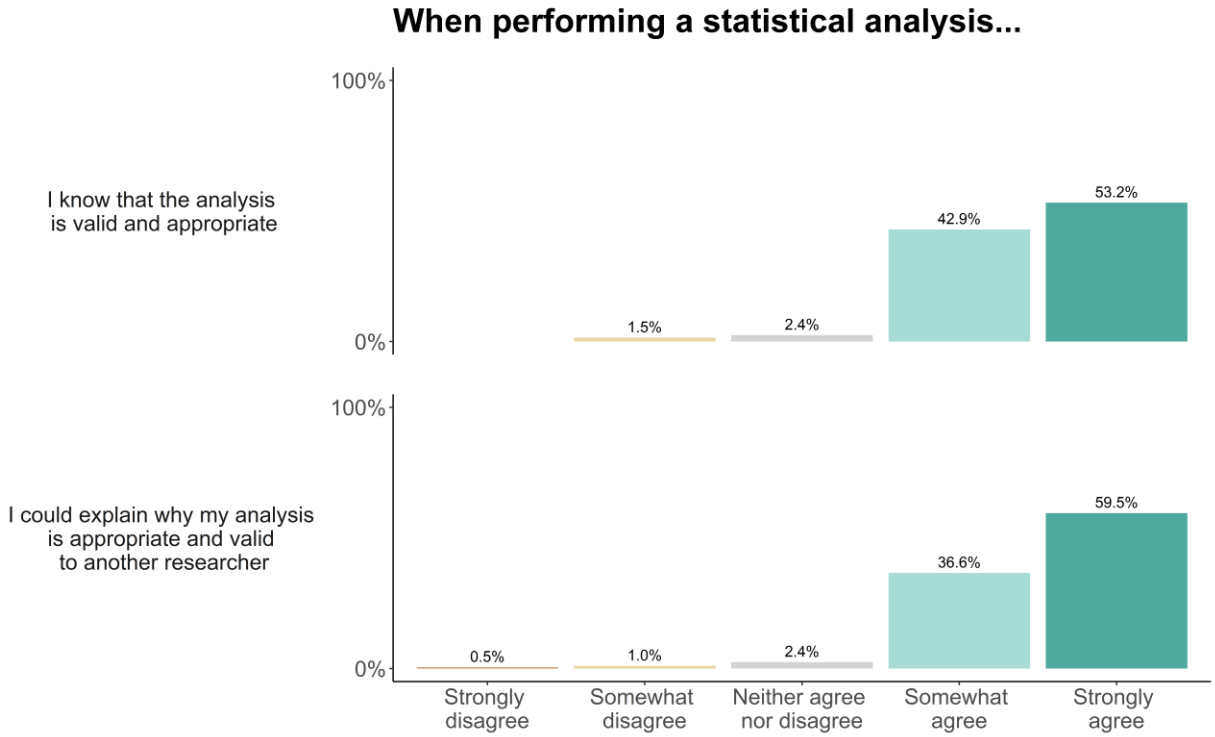


Figure 38: Animal cognition researchers' self-reported confidence in their own statistical analyses, N = 210.

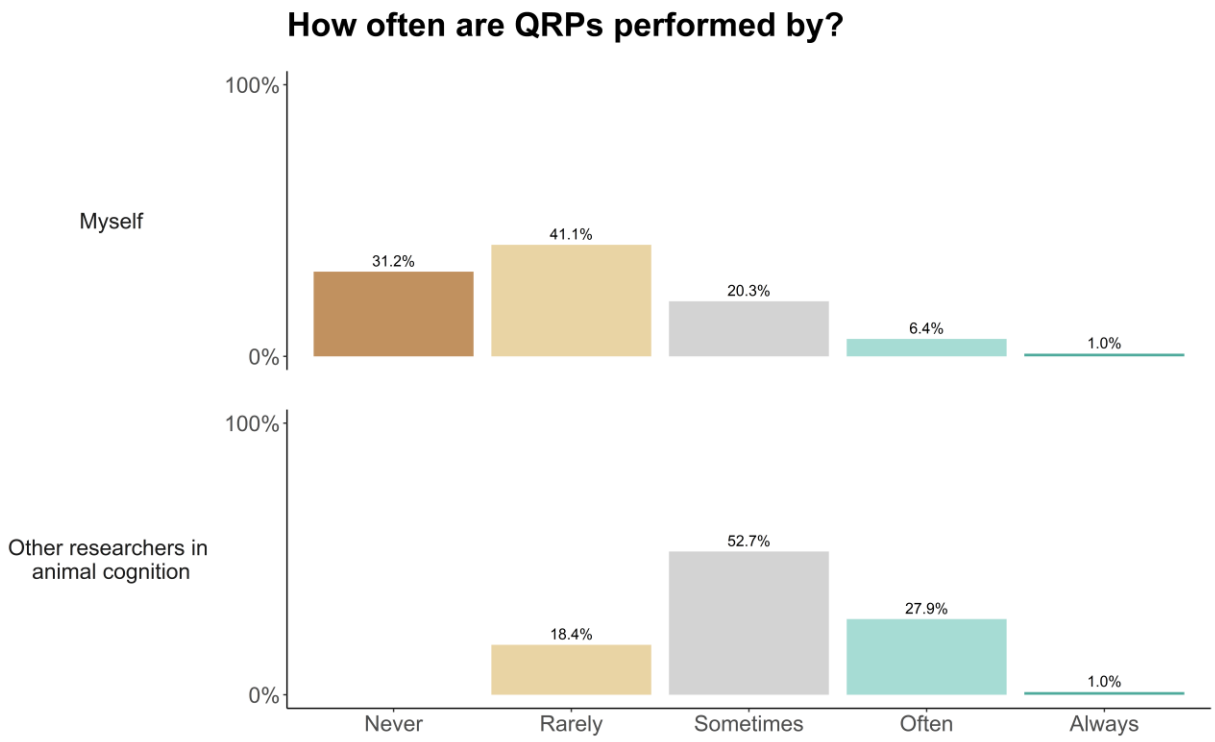


Figure 39: Animal cognition researchers' self-reported use of questionable research practices, and their estimated use of questionable research practices (QRPs) by other researchers in the field.

Table 26. Animal cognition researchers' comments about the use of statistics in the field. Four themes with a smaller number of responses are not shown (pre-registraion as a named solution (3), Bayesian statistics as a named solution (3), QRPs being less of a problem in animal cognition research (2), and the dangers of dichotomising research at p = 0.05 (2).

Do you have any comments about statistics in animal cognition?	N	Exemplars
Lack of training	15	*"I think that a lot of statistical mispractice also stems from missing knowledge/proficiency regarding statistical software and/or methods/tests. The majority of animal cognition researchers have a very sparse statistical education and are therefore self-taught. This can be a huge potential for errors."*
Complex statistics as a barrier	11	*"The increasing use of highly complex stats (e.g. Bayesian GLIM modelling) doesn't always help. I'm doubtful if most users can work out which variables are being treated as fixed effects in their analyses, for instances, and which of them should be. I certainly can't!"*

Do you have any comments about statistics in animal cognition?	N	Exemplars
		"This comment goes beyond animal cognition, but basically you need a second PhD in statistics to handle analyses these days. I think we are all doing our best, but there is only so much we can teach ourselves about statistics when we are also bogged down with our actual research, teaching, grants, etc."
A problem of low power or small sample research	9	*"Depending on the species, many problems stem from smaller sample sizes being treated in the same way as large samples."*
Incentive structure promotes questionable research practices	9	*"I also think current publishing requirements and standards are much to blame for these practices. Papers need to be short and concise, and it is more difficult to write a nice story when unexpected findings were not predicted. Journals also want the newer, exciting results more often than a simple, non-significant story. This doesn't mean these practices are acceptable or should be done, but not everyone can/will fight for ethical scientific standards when an easier solution is highly rewarded."*
Questionable research practices not necessarily bad	8	*"Sometimes it does happen that you conduct a study with very different intentions than the result you get. In hindsight it would have been a reasonable prediction, and framing it as such can help make a paper clearer." "In some cases I don't think there is anything wrong with this but there is a fine line."*
Discussed solutions	7	*"I worry about trendy "bandwagons" and fashions. Rather than prescribing particular approaches (e.g. we should all be Bayesians now) statistics should be reported clearly, transparently and in detail (e.g. I have no issue with people reporting p values if they want to, as long as they report effect sizes, associated errors and confidence intervals and visual representations of raw data)."*
An anecdote of QRPs	6	*"I have occasionally heard researchers pushing for collecting additional data to boost a trend, but it is difficult to estimate how common this practice is."*
Individual-level statistics important	5	*"At times, the search for population statistics obscures the attempts to understand individual variation. In other*

Do you have any comments about statistics in animal cognition?	N	Exemplars
		words, trying to forge a coherent analysis of many small NS may be more fruitful than statistics based on one large N."
Collaborating with statisticians	5	*"I find myself not so knowledgeable about new statistical techniques, so use a statistician."*

9.2.6.Replication

I asked researchers what proportion of replication studies they expect would be successful in their area of research, and to what extent their own and other areas of animal cognition would experience a replication crisis, if many of its studies were replicated. If 100 typical studies in their research area were replicated, researchers believed that 65% (IQR: 50% - 75%) would replicate successfully if the replication study tested a new sample of the same size with the same protocol as the original study. If these replication studies used sample sizes of 1000, researchers estimated that 72% would replicate successfully (IQR: 50% - 82%, Figure 40).

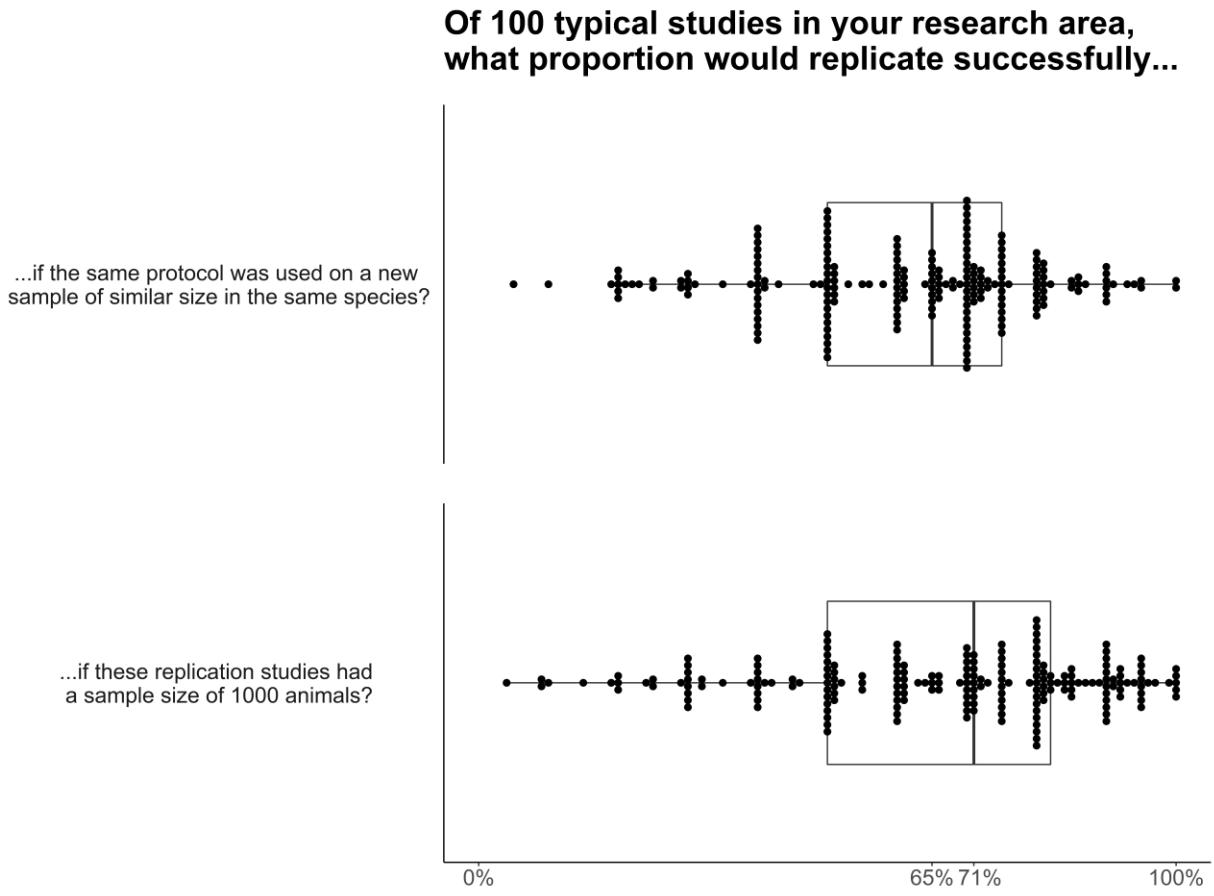


Figure 40: Animal cognition researchers' predictions of replication success in their field, $N_{same\ sample\ size}$ = 207, $N_{large\ sample}$ = 205.

Predominantly, researchers somewhat agreed (34.0%) or somewhat disagreed (30.1%) that their area of animal cognition research would experience a replication crisis if attempts to replicate most of its studies were conducted, and they either somewhat (43.7%) or strongly (29.3%) agreed that some other areas of animal cognition research would experience a replication crisis. Researchers tended to somewhat agree (38.0%), or neither agree nor disagree (31.2%) that they could identify which animal cognition studies would successfully replicate and which would not (Figure 41). When asked about the importance and prevalence of replication studies, researchers disagreed (50.7%) or strongly disagreed (20.1%) that enough replication studies were performed in their area of animal cognition research. These proprtions were matched when researchers were asked to consider replications in animal cognition research in general (disagree: 55.5%, strongly disagree: 23.6%). The vast majority of researchers agreed (34.8%) or strongly agreed (54.8%) that it is important that replication studies are performed in animal cognition

research (Figure 42). I received 64 free-text responses about replication in the field, with researchers most often highlighting various complexities and nuances of replication in animal cognition research Table 27.

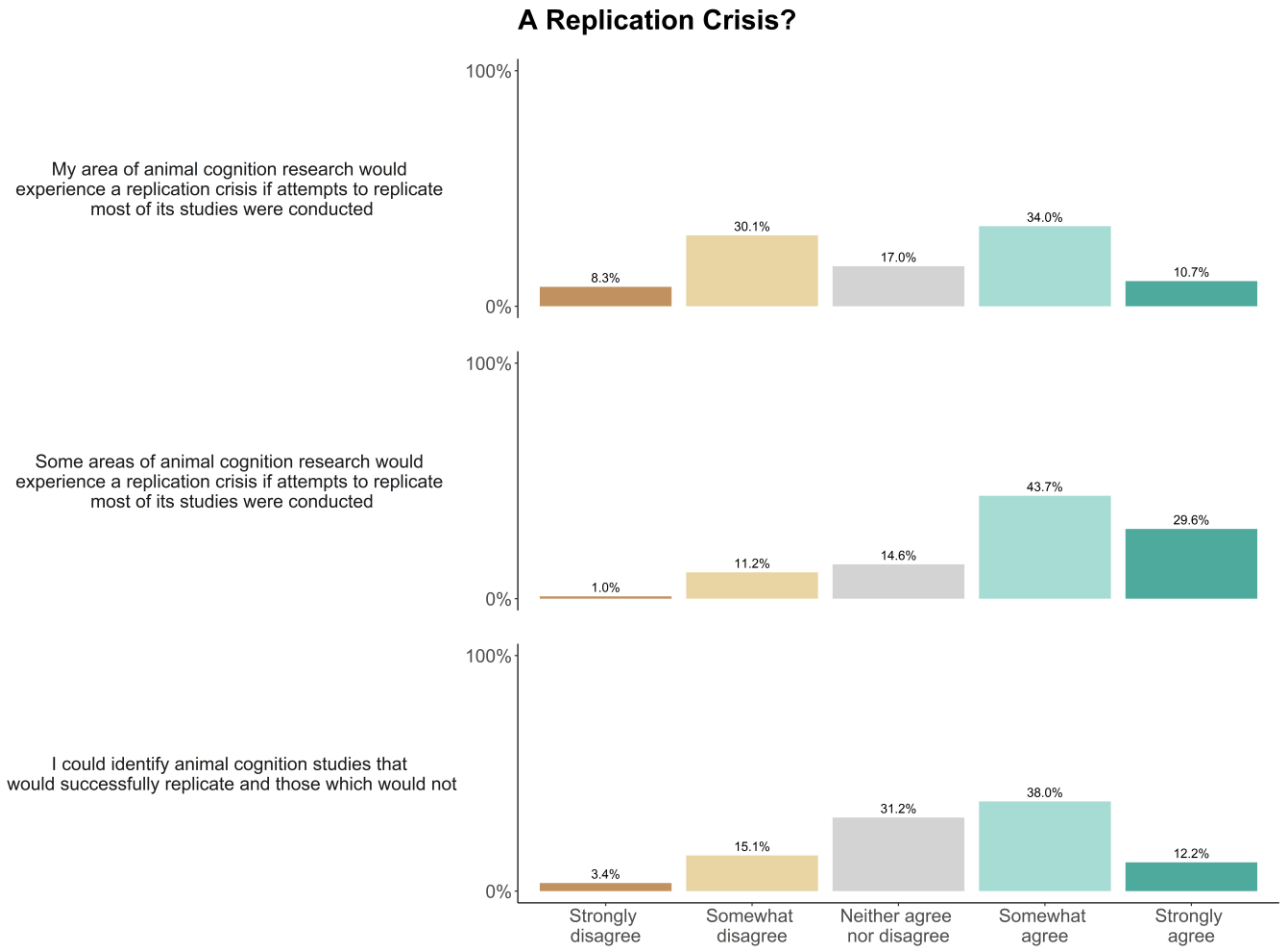


Figure 41: Animal cognition researchers' perceptions of a replication crisis in the discipline, and their ability to identity studies that would not replicate, N = 210.

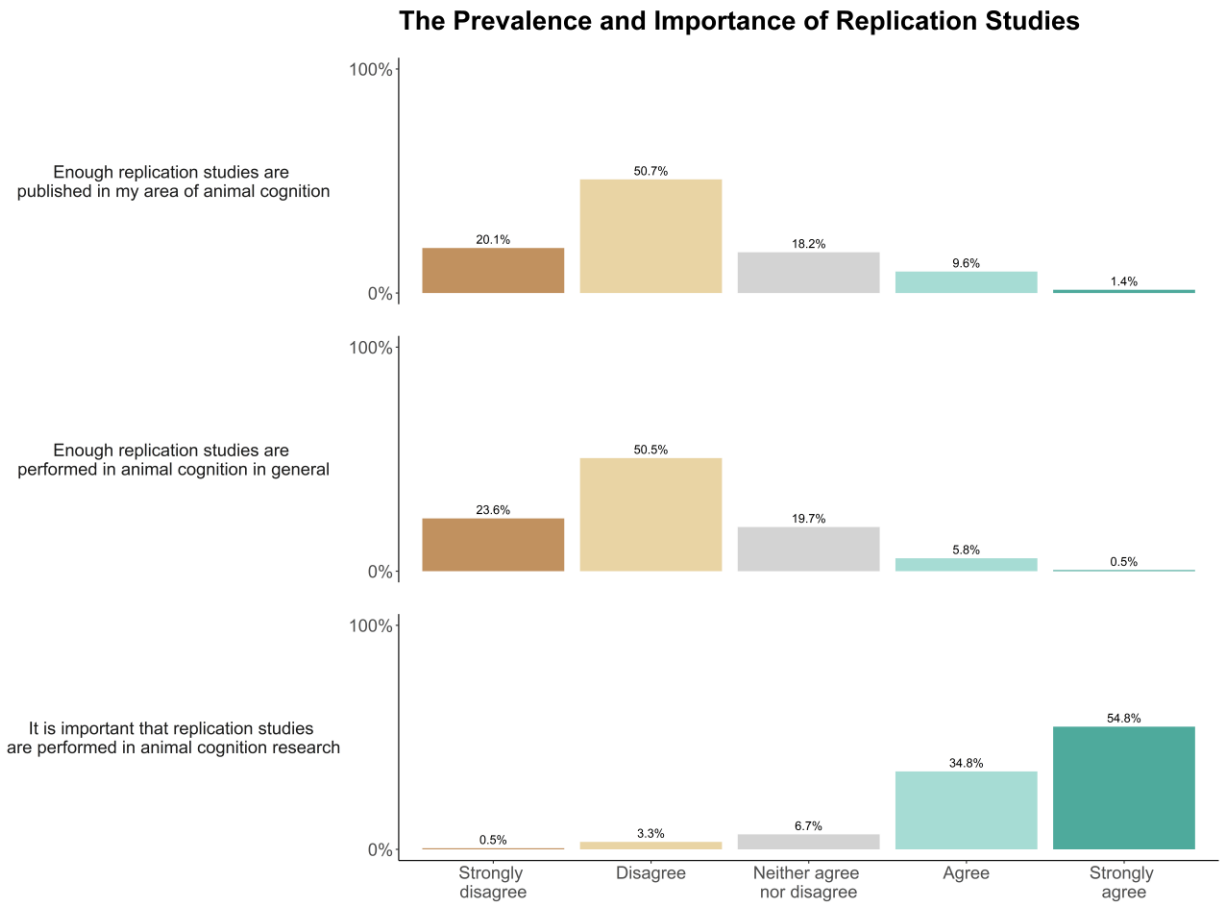


Figure 42: Animal cognition researchers' perceptions of the frequency and importance of replication studies in the discipline, N = 210.

Table 27: Animal cognition researchers' beliefs about replication in animal cognition research

Do you have any comments about replication in animal cognition?	N	Quotes
Complexity of replication	27	*"As results appear to be heavily influenced by all sorts of things - history and experiences of the animals, particularities of the facility, particularities of the group structure and (sub-)culture of the animals, the experiment paradigm details, reward distribution, training protocol... - replication is horribly difficult."*
Lack of Incentives	16	*"Journals are more and more looking for novel ideas and results, and unfortunately replication studies are seen as unimportant, unless they shockingly dismiss some big ideas."* *"Like other areas of research, the current publishing system values novelty and I believe this to be a major limitation that has discouraged replication in cognition research."* *"Funding to conduct replication studies is more difficult to obtain, than for novel studies"*
Importance of converging evidence	8	*"The term "replication" is not entirely straightforward. A strong replication is not always using the same protocol or the same stats and arrive at the same conclusion. I believe that a well-replicated result is something that shows to be correct when using a variety of methods and approaches and still arrive at a very similar conclusion. I think that this is true for several areas in animal cognition."* *"I think interpretation is a bigger issue than replication.* *Even under describing the methods (sometimes many important details are omitted) is a bigger issue.* *I DO think replication is important, but as we look across experiments, even though they are not exact replications, I think we can see the trends in what is likely a real effect and what may be something that could not be replicated. We should be training students how to look for these trends though."* *"Depends very much on the topic, and what is meant by replication. In controversial areas, such as episodic memory in animals, there has been numerous attempts to demonstrate or refute, but often with different species. This involves attempts to replicate a phenomenon, but not necessarily a particular study."*

Do you have any comments about replication in animal cognition?	N	Quotes
Issues with bias and validity more problematic than replicability issues	7	*"At least in my subfield of animal cognition the replicability of studies might actually be viewed as a negative because the assumptions and interpretations of the studies are fundamentally flawed. So the studies replicate, but researchers take those replications as additional evidence for the validity of their paradigm when it is not."*
Between area heterogeneity in replicability	5	*"Confidence in my own area of research has to do with the common practice of "embedded replications" of successful previous work in my work. My lack of confidence in some other cases has to do with demonstrations with sparse background literature/experimentation to back it up."*
Legal or ethical barriers to replication	3	*"One problem with replication in the context of animal work is the clash with the ethical pressure to minimise animal use."*

9.2.7.Belief

When reading papers in their own area of research, and other areas of animal cognition research, our sample reported often or sometimes agreeing with the authors' conclusions (own area: often: 58.4%, sometimes: 38.3%; other area: often: 58.5%, sometimes: 36.2%,

Figure 43). Researchers somewhat and strongly agreed that their beliefs about animals' cognition are affected by both scientific experiments (strongly agree: 55.7%, somewhat agree: 34.3%) and their day-to-day experience with animals (strongly agree: 31.9%, somewhat agree: 34.3%). When asked to choose between scientific experiments and experience with a slider response (with science at one extreme and experience at the other), researchers tended to say their beliefs were more driven by science, although a range of responses were observed (median: 31, IQR: 19 – 51, where 0 is exclusively based on science, and 100 exclusively based on experience, Figure 44). I received 42 free-text responses about beliefs in animal cognition, from which we identified 5 common themes. Table 28 outlines these themes and provides example quotes, and, although it did not fit one of my themes, I highlight another interesting quote below:

"I think you can almost always find a scientific paper to confirm your beliefs, and can find a way of justifying paying attention to that one, and ignoring one that might give different results. I don't mean this cynically -- but humans are very good at piecing together a plausible seeming story with limited evidence!

(We're good storytellers, and it can take a lot of evidence to dissuade someone from a good story!)"

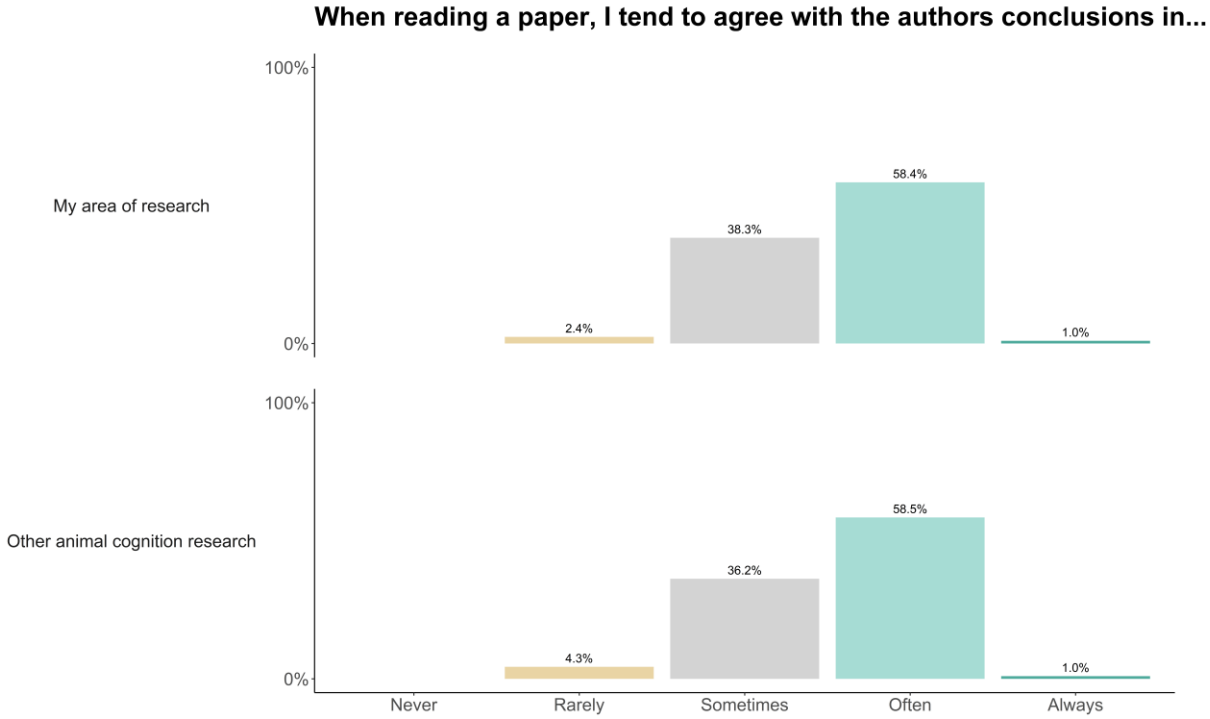


Figure 43: Animal cognition researchers' tendency to agree with the conclusions of papers in their own and other areas of research. N = 210.

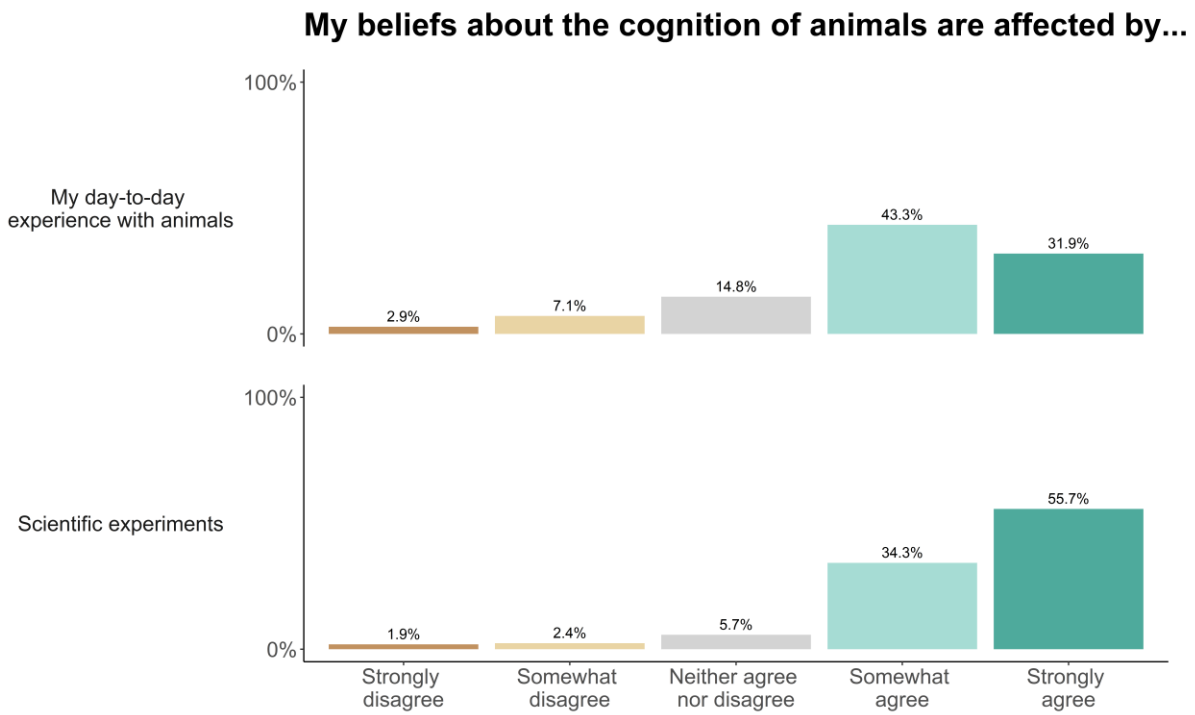


Figure 44: Animal cognition researchers' reports of the role of science and daily experience in shaping their beliefs about animals' cognition. N = 210.

Table 28: Animal cognition researchers' beliefs about the role of science and day-to-day experience in shaping their beliefs about the cognition of animals.

Do you have any comments about belief in animal cognition?	N	Quotes
Experience acts as a source of scientific hypotheses	11	*"Observations and experiences may give you hints about possible study questions. They leave you with impressions of animals' mind that require further digging into. Science, however, is absolutely vital to yield actual knowledge."*
Bias in scientific results prevents them impacting beliefs	10	*"I am hesitant to say that the results of scientific experiments affects my beliefs about animal cognition. This is mainly because I know that many studies are poorly executed, and it is the norm to make huge claims with no or limited data to back it up. Certain authors make careers out of their great skills at hyperbole, and I find this ethically unacceptable. On the other hand, there are authors that perform good science and don't make exaggerated claims. These studies I take seriously, and the work of such authors does indeed have the potential to affect my beliefs about animal cognition."*

Do you have any comments about belief in animal cognition?	N	Quotes
Science can answer questions experience can't	9	*"There are a lot of species studied...so even experts in the field could only obtain knowledge about that species by reading papers (for the most part)."*
The role of science and experience shaping belief varies depending on the topic	7	*"There are deep questions and shallow questions. Deep questions, like whether a crow has consciousness, can only be answered by a scientific theory of the concept. Shallow questions, like whether a dog has a memory of where a bone was buried, can be answered with empirical observations."*
Experience with animals can be valid data in itself, and/or necessary for producing valid data	5	*"What exists in the literature is relatively limited compared to the richness of experiences that working with animals regularly offers. Individual experiences, even if one-offs, can be very provocative indicators of cognitive potentials."* *"The definition of "cognition" is fuzzy and not recognized as "fuzzy," most students are taught to "operationalize" and to "standardize" their data, before they know enough about the natural behavior of the animals to be able to perform those types of procedures appropriately and it is in these procedures that bias inevitably and unwittingly enters into their research."*

9.2.8.Miscellaneous

Throughout the survey, five themes emerged across our survey blocks that our within-block coding did not identify. As such these themes were not those systematically extracted, but themes we believed came up across blocks and wanted to highlight. These were the role of theory in animal cognition, the need for an individual-level focus in research, academic incentives, the large amount of heterogeneity across animal cognition research, and the uncertainty surrounding the causes and implications of negative results. I provide representative quotes for each in Table 29. The most notable of these was a number of researchers who noted how their responses to individual questions would depend on who or what they were considering – something the Likert-type responses could not capture. For example, an individual may believe that on the whole most areas of animal cognition research are relatively unaffected by questionable research practices, but that they might be more common in certain sub-fields or research groups.

Table 29: Further topics commented on by animal cognition researchers that were not explicitly included in the survey.

Area	Quote
Theory	*"In my view a far bigger problem is poor theorizing [compared with replication]. A lack of formal theory (as exists in evolutionary biology) combined with "scala naturae" thinking, a lack of consideration of natural history and incentives to show that your study animal is "clever" or human-like are major problems for the field."*
Individual-level research	*"It is unfortunate that Single-Case experimental designs (single subject, single-organism, etc) are not used more often, which are known to (a) highlight replication and reproducibility, (b) avoid many hypothesis-testing issues (including, but not limited to those listed above), and (c) avoid many group-design limitations for behavioral research."*
Negative results	*"I usually never finish a study which I realize was misconstrued when I see the first behaviors of the animals. Oftentimes it is easy to arm chair-design a study which turns out to be impossible for practical and other reasons. This is not saying that I have not published finished studies with negative result. However, studies with negative results often needs additional controls to show it is a true negative; most often animal cognition studies are initially designed to control for that a potentially positive result is a true positive. There are many more ways for something to be negative than to be positive, therefor particular care must be given when publishing such data (negative or no results can often be the result of a bad design)."*
Incentives	*"In my opinion, the drive to publish 'exciting results' is driven by the expectations of funding bodies, and the general competitiveness of the academic system, that expect people to constantly produce ground-breaking new research. Not all research is or can be ground-breaking, but is a necessary part of research progress, such as proposals for methodological improvements. Such research deserves more support from the research community. Shifting the weight in expectations on researchers might reduce peoples' need to over-interpret borderline p values and report 'impactful' findings where there is really little to none."*
Heterogeneity	*"I don't agree with the question about my belief about other researchers tending to make weak or strong claims given their data. The answer should have been one of choosing the percent that I believe make stronger claims*

Area	Quote
	than warranted. Most researchers (70%) I believe make appropriate claims, but some (30%) do make stronger claims than warranted."

9.3. Discussion

This survey provides a picture of animal cognition researchers' beliefs about bias and scientific practice. From 1001 invitations, I received 210 completed surveys, from which I analysed data on a range of controversial topics and possible biases in animal cognition research. While it is likely that there was a self-selection bias in who completed our surveys, with researchers who have stronger feelings about bias in the field presumably being most likely to complete our survey, 210 completed surveys reflects a large number of recently active animal cognition researchers. Before discussing the individual survey topics, I want to outline what I believe data from surveys like these are useful for and what they are not. Specifically, I do not believe that these data are highly accurate quantitative estimates or representative data of what all animal cognition researchers' believe or how they behave. Rather, they must be interpreted considering the likely sampling biases in who participated in the survey and how their answers were limited by the way the questions were asked. Specifically, the strongest sampling bias is likely that the researchers who completed the survey, and especially those providing detailed free-text responses. These individuals are likely those who have thought most about some of the issues presented in the survey, and are potentially the most concerned about some of these issues (e.g., reliability) than researchers who did not complete the survey. This might mean that some of the quantitative estimates, e.g., perceptions of a replication crisis, might overestimate the "average" response of animal cognition researchers to this question, but equally might underestimate the concern about bias within their own results – if these researchers are more likely to e.g., adopt blinding strategies. Moreover, it is likely the manner in which the questions were phrased had some unavoidable influence and priming effects on the participants – while no attempts were deliberately made to bias participant answers in a particular direction, it should be appreciated that different phrasings of the questions would likely return slightly different responses (Fiedler & Schwarz, 2016). Nevertheless, each individual response that I received reflects the opinion of a particular animal cognition researcher, and thus are inherently meaningful pieces of data, with detailed full-text responses available at osf.io/6j7kp.

9.3.1. Bias

Overall, researchers were wary of bias across animal cognition research. Researchers often agreed, or neither agreed nor disagreed, that the results and theories across animal cognition are strongly affected by researchers' biases. For example, some researchers' qualitative responses suggested that they believe bias not to be uniform across the field, instead reporting that certain topics and researchers may be more likely to be affected by bias than others. Similar to other survey studies of scientific bias, participants were generally more concerned about bias in others' research than their own (Fraser et al., 2018; John et al., 2012), although there were exceptions, often being both very conscious about the possibility of bias in their own and others' work. This was especially pronounced for experimenter bias, where researchers did not appear especially concerned that they might be biasing their own results, and were, on average, confident they could perform fair tests of animal cognition. This somewhat conflicts with primary data suggesting that experimenter effects can have a large influence on animal behaviour, and that blinding procedures are rarely reported (Bohlen et al., 2014; Lit et al., 2011; see discussion of blinding in Chapter 6). This confidence in avoiding experimenter effects might reflect an overrepresentation of researchers in our survey who take steps such as blinding to minimise these effects in their research, or who believe their experiments should be unaffected (e.g., by not being in contact with animals during testing due to using touchscreen apparatus). However, I also received some strong responses from researchers who fervently believed that researchers always hope for particular results and thus should always be concerned that they might be biasing their results, and several researchers noted how bias can be embedded in research programmes even before data collection begins.

My survey results also provide direct evidence of publication bias in animal cognition research, self-reported by active researchers in the field. The median percentage of studies researchers reported publishing was 80%, although over 10% researchers reported publishing less than 50% of their studies. These figures may underestimate the prevalence of publication bias both within my sample and in animal cognition more generally. Within my sample, the figures may be an underestimate as published findings are likely easier to recall for participants while they were completing the survey (i.e., an availability bias (Tversky & Kahneman, 1973)). In animal cognition more broadly, the figures may be an underestimate if my participants were more likely to publish negative results than the average animal cognition researcher. While researchers reported a journal or reviewer enforced publication bias against negative results or against results not in line with "preferred" theories, many researchers also reported not attempting to publish studies with difficult to interpret results, or those that had flaws in the experimental design or were otherwise perceived to be low quality. Notably, this decision not to publish was often the researcher's own, with a lack of time or incentives often cited as the limiting factor. Combining participants' quantitative and

qualitative responses suggests that across most areas of animal cognition research, many studies have been performed but not published, which is in line with the statistical markers I presented in Chapters 5, 6 and 7.

9.3.2.Morgan's canon

Over 70% of the sample somewhat or strongly agreed that Morgan's canon is important to use when interpreting the results of animal cognition experiments. Superficially, this contrasts with a large body of literature criticising the canon on the grounds that there is no reason to privilege "simpler" or "lower" explanations of animal cognition over more "complicated" or "higher" explanations (Andrews, 2020; Bausman & Halina, 2018; Buckner, 2013; Sober, 2005). However, participants qualitative responses revealed a more nuanced picture: Many of those who also provided free-text responses, a) recognised the inherent ambiguity and multiple interpretations of Morgan's canon, and, b) cautioned against a blind application of Morgan's canon. Of those who defended the canon, most defended a particular principle associated with it (e.g., parsimony and phylogeny), rather than the canon itself. Evidently, Morgan's canon and related concepts elicit a plurality of opinions. Because of the variety of interpretations and justifications for invoking the canon, or e.g., parsimony, arguments should not likely be evaluated based on the authority of these principles alone – because researchers might understand them differently. Rather, researchers should strive to make the assumptions and justifications for favouring one hypothesis over another explicitly – something that could be achieved through formal modelling (although, see "Theory and modelling" section in discussion).

9.3.3.Replication

Over 70% of the sample agreed or strongly agreed that some areas of animal cognition could experience a replication crisis, and, in my sample, slightly more researchers agreed (44.7%) than disagreed (38.4%) that their own area of research would experience a replication crisis, if attempts to replicate its studies were performed. This suggests a large degree of skepticism about the robustness of research findings in some areas of animal cognition research, or of the ability of replication studies to repeatedly identify certain effects. However, such skepticism is common across sciences, with 52% of 1576 researchers surveyed across fields including biology, chemistry and physics, reporting that there was a "significant" reproducibility crisis in their field.

Researchers near unanimously agreed that replications were important, and not performed frequently enough (Figure 42), mirroring the view of ecology and evolution researchers (H. Fraser et al., 2020). And A smaller number of researchers noted that replication studies may be less important than seeking convergent evidence of phenomena. These views echo wider discussions about the role

of direct and conceptual replications in psychology, with conceptual replications being essential to provide robust evidence of general psychological effects (see e.g., Crandall & Sherman, 2016).

9.3.4.Belief

Researchers reported that their beliefs about animal cognition are influenced by both the results of scientific experiments and their own personal experience with animals. Typically, researchers viewed science and experience as synergistic, with experience often cited as the source of scientific hypotheses, and necessary for designing good experiments. A smaller number of researchers also endorsed every-day knowledge as a valid source of data that could be seen as equally strong as some scientific data (Fraser et al., 2017), although researchers often noted that the role of science and experience depended on the question at hand – there are some, often trivial, questions that can be answered readily through experience, yet many researchers reported that some knowledge can only be accessed through systematic scientific study. Finally, researchers noted that for many species that they have no experience, rely on the scientific literature to form their beliefs, which requires them to trust the findings of their colleagues.

9.3.5.Miscellaneous

While my survey focused on five blocks of questions that I was particularly interested in at that stage of my PhD, oftentimes researchers' free-text responses went beyond these questions and highlighted specific issues that were not directly solicited by the survey. For example, a researcher offering reservations about the press coverage of animal cognition research, or species biases in what is tested and interpreted, as well as biases based on the location of where research is conducted. We encourage the reader to view the full database of open-text responses to make the most use of these low-frequency data from this survey (osf.io/6j7kp). However, there were five themes that I interpreted that went beyond my initial survey aims. These were theory, individual-level research, incentives, heterogeneity and interpreting negative results, and these are discussed in the "Ways Forward for Animal Cognition Research" section of Chapter 10.

9.4. Summary

This survey provided a snapshot of animal cognition scientists' beliefs about bias, replicability and practices in animal cognition research. Animal cognition scientists predicted replicability issues in the field and were generally wary of a range of biases affecting the research process, although more so in others' work than their own. The data give credence to the arguments made in Chapters 2, 3 and 4 of this thesis – that areas animal cognition research may contain many results that will struggle to replicate, and be heavily influenced by theoretical biases, in part due to academic incentives. The also

survey provided direct evidence for a publication bias affecting the field, in-line with findings from Chapters 5, 6 and 7: researchers self-reported publishing a median of 80% of their studies, however, there was a considerable variation in their responses. Publication bias seemed to be against negative, difficult to interpret or poorly designed research, and was both reported as self-enforced (i.e., the article was never written or submitted), and journal enforced. Researchers also perceived a journal- and reviewer-enforced publication bias against results contra to established theories and reviewers' preferences. On the whole, participants displayed a range of opinions concerning bias and replicability, largely mirroring the debates of the wider scientific community when considering reliability of scientific results. These views included advocating for incentive reform and replications, and improving statistical inference, but also stressing the importance of developing theory and seeking converging evidence for theories.

10. Chapter 10: Discussion

10.1. Overview of thesis findings

This thesis started by introducing the replication crisis (Chapter 1) and asking why animal cognition research had not yet experienced these issues (Chapter 2). Answering this question led me to explore what exactly constitutes a replication and how this relates to theory testing, applied to animal cognition research (Chapter 3). However, I found that low replicability might be just one symptom of deeper scientific and incentive issues in animal cognition (Chapter 4), and I then developed methods to quantitatively assess a range of biases and critically synthesise evidence in animal cognition research (Chapters 5 to 9). As such, the thesis provided four general contributions to the study of animal cognition. First, it performed a long overdue integration of the replication crisis literature and animal cognition research, concluding that many areas of animal cognition research likely contain many difficult to replicate findings (Chapters 1, 2, 3 and 4). Second, it made a novel argument about the nature and causes of certain biases in animal cognition research, especially in areas claiming evidence of higher-order cognition in animals (Chapter 4). Third, it developed, implemented, and critiqued secondary data analysis techniques capable of critically synthesising evidence and assessing research practices in animal cognition research (Chapters 5, 6, 7 and 8). And fourth, it surveyed researchers' opinions on the topics of this thesis, making public an important, but often tacit, knowledge-base on bias in animal cognition research (Chapter 9).

In addition to these four general contributions, the thesis made three specific contributions. First, Chapter 5 mapped the types of statistical inferences used in animal physical cognition research and provided some evidence of publication bias through a novel claim categorisation method – having many coders assess whether claims were "positive" or not. Second, Chapter 6 mapped and critically assessed the near entirety of corvid social cognition literature. It again found evidence of publication bias, and a concerningly low levels of reported experimenter blinding[14], and data and code availability. And third, Chapter 7 described and assessed how researchers make inferences from results that are not statistically significant – an issue which my survey sample repeatedly highlighted as a challenge for animal cognition research (Chapter 9).

10.2. Is animal cognition research in a replication crisis?

In Chapter 2, I outlined why I thought many areas of animal cognition research likely contain many difficult to replicate findings, but also how some areas – typically those employing many trial designs

[14] Of course there are some tasks for which blinding might be difficult to achieve, e.g. object choice tasks in which an animal is required to respond to another agent's cue (blinding could be achieved by using a naïve or asocial agent, but this might change the fundamental nature of the task).

– are likely more robust. This argument was made at the start of my PhD, and over the course of the PhD, the predictions have largely borne out. Some replication failures have been published (e.g. in avian cognition Amodio et al., 2021; Crosby, 2019; O'Neill et al., 2021; Soler et al., 2020), alongside an increasing number of newly published negative results (e.g., Amodio, Brea, et al., 2021; Brecht et al., 2018). Per the predictions of Chapter 2, these negative results and replication failures have struggled to pinpoint the cause of these negative results to low power, idiosyncratic samples or false positive original findings. Encouragingly, the field is displaying increasing attention and caution about replication across the literature, through: special issues on the topic (Brecht et al., 2021); specific articles about replication (Beran, 2018; Boyle, 2021; Dacey, 2020; Halina, 2021; Khan & Wascher, 2021; Shaw et al., 2021; Stevens, 2017; Tecwyn, 2021); general articles about animal cognition research (Bastos & Taylor, 2020; Krasheninnikova et al., 2020; Schubiger et al., 2020); in reviews of individual research topics (Colbourne et al., 2021); in PhD theses (Amodio, 2020; Crosby, 2019; Schubiger, 2019); and within individual experiments (Bohn et al., 2021; Szabó et al., 2017).

Nevertheless, the question of whether animal cognition research is in a replication crisis has not been resolved, and nor should it have been. The field is too heterogenous to be labelled as such. While it could be viewed a "crisis" in the sense that most areas have not demonstrated their reliability, the reality is we still know very little about the robustness of many areas of animal cognition research. Replication, and up-to-date critical systematic reviews, are needed for each individual research area within animal cognition that we wish to understand the quality of evidence for. However, performing and interpreting these replication studies and reviews is not straightforward (see Chapters 2, 3, 5, 6 and 8), and likely resource intensive, which may be especially problematic for areas of animal cognition research facing reduced funding.

10.3. Bias, incentives, and animal cognition research

In Chapter 4, I argued that academic incentives have selected for poor research practices across much of animal cognition research. Combined with normalised research and publication practices that produce many false positive or overstated results, many claims in animal cognition can be traced back to these incentives, and not the actual cognitive abilities of the animals. I argued that this bias most clearly manifests in researchers using null-significance hypothesis tests to claim the presence of complex cognitive abilities in animals. Such research programme may have their overarching conclusions set before data collection begins, and the research acts as a negotiation as to how these conclusions will be reached, and not whether. However, I also noted how the same incentives likely affect animal cognition research throughout – including the skeptics.

My argument relied predominantly on indirect evidence and by analogies. This was largely because of the paucity of data on replication, statistical power and false positive rates in animal cognition research itself. Such a lack of data likely limits the ability of my argument to convince those skeptical of it, although many respondents of the survey study in Chapter 9 shared similar reservations. Gathering evidence capable of assessing factors such as publication bias, statistical power and replicability will be key to assessing where the argument in Chapter 4 applies and where it does not. In the interim, it is up to individual researchers to decide what they believe about the likely strength of different areas of animal cognition research. To take the example of research programmes that have consistently produced evidence supporting the presence of more and more complex cognitive abilities in their study species. My opinion is that because, a) there are clear incentives to produce such findings, and b) biasing factors such as publication bias and alpha inflation have been normalised parts of scientific research, the onus of evidence should be on these research programme to demonstrate that their results are reliable and valid. This either requires a suite of registered, "best-practice" direct and conceptual replication studies, or strong retrospective secondary data analysis to explain why the published literature is at low risk of biasing factors.

In Chapters 5, 6 and 7, I attempted to develop such methods capable of detecting publication bias and assessing evidential strength in animal cognition research. None of the findings of these Chapters were inconsistent with the argument I made in Chapter 4, and they would have likely produced different results if the areas of animal cognition research I focused on (animal physical cognition research and corvid social cognition research) contained exclusively reliable research. Further developing and implementing such secondary data analysis projects should be a goal for individual research themes within animal cognition, and the likely utility of and barriers to this were discussed in Chapter 8.

10.4. Ways forward for animal cognition research

While many findings might be unreliable in animal cognition research, the field may be too resource (and incentive) constrained to replicate all its key findings. In addition, widespread disagreement about the validity of previous research also questions the utility of replicating it. For the remainder of the discussion, I therefore focus on how the lessons of this thesis can be used to prospectively increase the quality of research in animal cognition.

10.4.1. Overcoming low power

Low power is an issue across some, but not all, animal cognition research (Chapter 2). For areas of animal cognition that do not know whether low power is an issue, they should prioritise calculating

the statistical power of their most common designs to detect theoretically interesting effect sizes. Resources are now available for conducting power analyses, and for specifying effect sizes of interest (e.g., Lakens, 2021). If low power is deemed to be an issue, researchers could consider increasing training and trail number in their studies, or recruiting more animals:

10.4.1.1. Increase training and the number of trials

Increasing trial number is likely the most efficient method of increasing power in animal cognition experiences (see Rouder & Haaf, 2018 for a discussing of increasing trail number vs sample size). Moreover, given that psychological effects usually occur within individual animals (e.g., learning occurs *within* an individual animal, causal reasoning would occur *within* an individual animal; Craig & Abramson, 2018; Skinner, 1956) a clear case can be made that researchers should design their experiments with the statistical power to detect meaningful effects within individual animals (Smith & Little, 2018). This has the twofold benefit of increasing the reliability of research findings (high power at the individual level entails high power at the group level), but also of being able to quantify and describe meaningful individual differences in behaviour. Similarly, increasing the amount of training animals receive, such that they are familiar and comfortable with apparatuses and testing procedures before the critical test will reduce the amount of "noise" in a dataset. The training pulls all individuals towards their theoretical maximum, increasing statistical power and the relevance of the collected data to the theory in question (Schank & Koehnle, 2009; Smith & Little, 2018), and can increase the validity of between-group comparisons when the groups have markedly different learning histories (Leavens et al., 2019).

10.4.1.2. Recruit more animals

Some designs cannot easily employ extra trials or increased training, for example those that rely on an animal's first exposure to a stimulus. Here, researchers should consider increasing the number of animals they recruit. Pre-study power analyses can help the researcher understand what sample sizes would be needed to detect certain effect sizes. Unfortunately there may only be minimal benefits of increasing sample sizes by 5 or so animals, which might be the most that can be recruited at any site (alternative approaches would be to increase the salience of treatments/interventions). Multi-laboratory studies hold some promise to resolve these issues, such as the ManyPrimates, ManyBirds and ManyDogs collaborations (Lambert et al., 2021; Many Primates, Altschul, Beran, Bohn, Call, et al., 2019; Many Primates, Altschul, Beran, Bohn, Caspar, et al., 2019; ManyPrimates et al., 2021). In principle, these are the types of studies that can be capable of quantifying meaningful between-species differences in behaviour, and the ManyPrimates collaboration is currently a strong example of how to perform such a project. However, multi-laboratory collaborations in animal

cognition research should also be wary of, i) overinterpreting "species differences" in behaviour, if single species are only represented at single sites (Chapter 3), and, ii) the increasing difficulty of independent criticism of their projects as collaborator number increases (Danchev et al., 2019).

10.4.2. Overcoming publication bias

There is no good reason for publication bias to continue to exist in animal cognition research. A field with publication bias against negative results risks becoming dominated by over-estimated and false effects (Chapter 2) and a need for close replication studies. Without publication bias, the field may instead focus on performing more conceptual replication and stronger tests of theory (see e.g. Halina, 2021, for a discussion about replication in comparative psychology and the effects of publication bias).

Fortunately, there are a range of low-barrier options to publish without biases against certain results in animal cognition. Venues such as *Animal Behavior and Cognition* frequently publish peer-reviewed negative results with open access and no author processing charges (e.g., in the latest issue at the time of writing: Eckert et al., 2021; Troisi et al., 2021; Wilson et al., 2021). And pre-prints offer a zero-barrier option to researchers who wish to publish any findings, which may be suitable for archiving experimental failures that researchers would not otherwise submit to journals. Finally, the registered report journal format offers pre-data collection peer-review and in-principle acceptance of articles independent of which results manifest – which has the added benefit of being able to improve study design before data collection begins (Chambers, 2013; Motes-Rodrigo et al., 2021; Tysall et al., 2020; Vonk & Krause, 2018).

While scientific incentives might still discourage some researchers from publishing some negative results (see Chapter 9, and section "overcoming perverse incentives" in this chapter), there are now minimal venue-based barriers for willing researchers to publish negative results. A longer-term goal for animal cognition research may be to emulate research areas with public registration of studies – which is the norm in clinical trials research (e.g. https://clinicaltrials.gov/) and soon to be law in the EU (Rasendriya, 2021). Such registration would allow the full extent and consequences of publication bias to be traced – although other publication-related biases, such as citation biases, could remain.

10.4.3. Overcoming ambiguity through theory and modelling

Chapter 4 highlighted how ambiguity is a major barrier to evidence assessment in animal cognition. Without knowing what, exactly, a researcher or article is claiming, assessing the strength of these claims are difficult to assess. One method that has been proposed to overcome these issues is through using "stronger theory" and/or formal models (Allen, 2014).

A main source of ambiguity in animal cognition is testing vague, verbal hypotheses that are only loosely connected to the data the experimenters collect (Chapter 4, Chapter 9). Data from these hypothesis tests can be interpreted in almost any way an individual chooses. In contrast, formal theories, be they logical, computational or mathematical, can have a string of benefits. For example, they might increase the precision and communication of hypotheses, make clear predictions, and offer the ability to simulate effects (see Farrell & Lewandowsky, 2010; Guest & Martin, 2020; Maatman, 2021 for discussion). In animal cognition research, evolutionary theory (Vonk & Shackelford, 2012), and learning theory (Dickinson, 2012; Skinner, 1976), are two possible sources of strong theory to ground research programmes in, and tools and tutorials for using theories like this in study design and analysis are increasingly available (Cinar et al., 2020; Jonsson et al., 2021).

Nevertheless, the extent to which formal models can effectively be generated for all research lines of interest is unclear, and nor do they guarantee easy to interpret results (Smith et al., 2012, 2016). This uncertainty can be illustrated by the example of animal theory of mind research. Whether animals represent others' mental states animals is an interesting question, and some strong modelling and associated empirical work on theory-of-mind in animals has been performed (e.g., Thom & Clayton, 2013; van der Vaart et al., 2012; van der Vaart & Hemelrijk, 2014). However, such models have not become mainstream in the study of theory of mind in animals. I believe this is because such models are i) asymmetric, ii) specific, and iii) require expertise. They are asymmetric as the models are often built upon abilities often seen as alternative hypotheses by researchers (e.g., associative learning, stress), they are specific to individual cases (e.g., corvid re-caching), and require expertise that many animal cognition researchers may not have the time to develop in the current academic climate. To what extent animal cognition researchers wish to use modelling returns to Beach's (1950) original dilemma – how can animal cognition best focus its limited resources? Is it on a range of questions with little independent validation and surface level experiments, on a few questions with intensive research, or some mixture or middle ground? Here, there is no clear solution, and the outcome will likely be driven by wider scientific incentives.

10.4.4. Overcoming perverse incentives

Changing scientific incentives requires action at the level of the individual and laboratory (Yarkoni, 2018), organization (Nosek et al., 2012) and even society (Amann, 2003; Lazebnik, 2018; Stengers, 2000; Stengers & Muecke, 2018). Psychology's reform movement has ensured that the knowledge, training and infrastructure for researchers to perform more reliable research are now widely available – whether that be in the form of articles, online training courses and tutorials (Herrera-Bennett et al., 2020), journal clubs (Orben, 2019), novel publication outlets and formats (Chambers, 2013; Vonk &

Krause, 2018), or wider infrastructure, such as that provided by the Centre for Open Science (www.cos.io). Nevertheless, if scientific incentive and funding structures continue to select for compelling narratives, oversold findings, and ground-breaking results researchers may continue to be disadvantaged by being more rigorous with their work (Higginson & Munafò, 2016; Smaldino & McElreath, 2016). Hence, top-down initiatives will be crucial in changing the scientific method of biological and psychological researchers – animal cognition researchers included. Such initiatives are gaining traction, such as the Declaration on Research Assessment (DORA: https://sfdora.org/), and funding mandates for data sharing (e.g., BBSRC 2007). Evidently, animal cognition researchers can only play a small role in such policy-making, and the role individual researchers can play in improving research practices and incentives in animal cognition will vary between personal circumstances.

10.4.5. Recalibrating aims, expectations, and questions

The window of possible replication crisis in animal cognition research provides an opportunity to better understand the field as a science, including its aims and expectations. All else being equal, demands to improve reliability and validity require more resources to be focused on fewer topics, and stakeholders in animal cognition research must consider whether this is desirable or not. Is it better to gather lots of data on a single topic, which might lead to strong answers about the nature of animal minds (but again it might not), or is it better to produce surface-level data on a wider range of topics, in the knowledge that this will rarely produce consensus answers about animal minds, but nevertheless data that can facilitate discussion between interested parties? Heterogeneity aside, I believe the latter scenario is probable if animal cognition continues with business-as-usual, but with the reduction of publication bias and increased statistical power. The former scenario – generating more belief constraining evidence – would require more formal modelling, more causal analysis and more test development (Fiedler et al., 2021; Flake & Fried, 2019; Rohrer, 2018). The heterogeneity of animal cognition research, however, means that different areas of the field might settle on a different balance of these approaches. Indeed, effective scientific research is likely present across some areas of animal cognition research, but for an outsider pinpointing which areas is currently difficult (Chapter 4).

Importantly, animal cognition researchers may wish to focus on greater uncertainty communication regarding their results. When I started this PhD the degree to which data undetermined claims and theories in animal cognition (Boyle, 2021) was not clear to me from the published literature. The continued publication of general claims from single or few experiment studies, and synthesis of this research without considering factors such as publication bias, suggests that this is still an issue. This is especially the case when the results of animal cognition research filters

into popular media: the science of animal cognition that the public consumes looks very different to the science of animal cognition performed in the laboratory and the field. Whether this is a desirable or even a somewhat necessary situation for animal cognition research to be in is a question worth debating in the literature.

Finally, a promising alternative is to shift more of animal cognition's research effort away from hypothesis testing of cognitive theories (see Scheel et al., 2021), and instead increase the amount of descriptive and exploratory research. Such descriptive research may then lead to hypothesis testing with a tighter derivational chain between statistical model and substantive claim. Such data could provide the basis for interesting behaviours to be identified, and subsequent cognitive theories to be proposed and tested. But importantly, the results of these tests should be contextualised relative to the body of descriptive work that had been conducted prior to it. This stands in contrast to the current landscape of top-down motivated research trends seen today (de Waal & Ferrari, 2010; Eaton et al., 2018; Vonk, 2021). But, if animal cognition research wants to retain the diversity of species and topics Beach (1950) pined for, alongside having confidence in the data it produces, then well-collected and well-reported descriptive research could be an informative and crucial step before complex cognitive hypotheses are tested.

10.5. Concluding remarks

Understanding animal minds, that are in principle unobservable, is a challenging task. If conducted within a scientific incentive structure that selects for ambitious claims and against rigorous science, the task risks being insurmountable. This thesis has reflected my attempts to grapple with these issues in animal cognition research, a process that left me with little confidence in much of the published literature on animal cognition, especially in the case of research programmes claiming to demonstrate higher-order cognition in animals. Importantly, this lack of confidence is not a disbelief in the cognitive abilities of animals themselves – I am largely agnostic to most of these questions – but is a disbelief in the strength of evidence that is portrayed across the literature. Whichever path animal cognition research takes in the future, Beach's (1950) trade-off between rigour and breadth will be a constant point of debate. While conducting this debate, animal cognition researchers should strive to conduct their research in a manner that facilitates transparent, cumulative, and critical evidence synthesis, and one that avoids the pitfalls of academic systems and journalism that promote sensationalism.

References

Aad, G., Abbott, B., Abbott, D. C., Abed Abud, A., Abeling, K., Abhayasinghe, D. K., Abidi, S. H., AbouZeid, O. S., Abraham, N. L., Abramowicz, H., Abreu, H., Abulaiti, Y., Acharya, B. S., Achkar, B., Adam, L., Adam Bourdarios, C., Adamczyk, L., Adamek, L., Noel, D., … The ATLAS collaboration. (2021). Search for supersymmetry in events with four or more charged leptons in 139 fb^{-1} of $\sqrt{s}$ = 13 TeV *pp* collisions with the ATLAS detector. *Journal of High Energy Physics*, *2021*, 167. https://doi.org/10.1007/JHEP07(2021)167

Aczel, B., Palfi, B., Szollosi, A., Kovacs, M., Szaszi, B., Szecsi, P., Zrubka, M., Gronau, Q. F., van den Bergh, D., & Wagenmakers, E.-J. (2018). Quantifying support for the null hypothesis in psychology: an empirical investigation. *Advances in Methods and Practices in Psychological Science*, *1*, 357–366. https://doi.org/10.1177/2515245918773742

Adams, C. D., & Dickinson, A. (1981). Instrumental responding following reinforcer devaluation. *The Quarterly Journal of Experimental Psychology Section B*, *33*, 109–121. https://doi.org/10.1080/14640748108400816

Adriaense, J. E. C., Martin, J. S., Schiestl, M., Lamm, C., & Bugnyar, T. (2019). Negative emotional contagion and cognitive bias in common ravens (*Corvus corax*). *Proceedings of the National Academy of Sciences*, 201817066. https://doi.org/10.1073/pnas.1817066116

Agnoli, F., Wicherts, J. M., Veldkamp, C. L. S., Albiero, P., & Cubelli, R. (2017). Questionable research practices among Italian research psychologists. *PLOS ONE*, *12*, e0172792. https://doi.org/10.1371/journal.pone.0172792

Alberts, B. (2013). Impact factor distortions. *Science*. https://www.science.org/doi/abs/10.1126/science.1240319

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

Allen, C., & Bekoff, M. (1997). *Species of mind: The philosophy and biology of cognitive ethology*. MIT Press.

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

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

Amodio, P. (2020). *Convergent cognitive evolution: What can be learnt from comparisons with corvids and cephalopods?* https://www.repository.cam.ac.uk/handle/1810/305051

Amodio, P., Brea, J., Farrar, B. G., Ostojić, L., & Clayton, N. S. (2021). Testing two competing hypotheses for Eurasian jays' caching for the future. *Scientific Reports*, *11*, 835. https://doi.org/10.1038/s41598-020-80515-7

Amodio, P., Farrar, B. G., Krupenye, C., Ostojic, L., & Clayton, N. S. (2021). Little evidence that Eurasian jays protect their caches by responding to cues about a conspecific's desire and visual perspective. *ELife*, *10*, e69647. https://doi.org/10.7554/eLife.69647

Amodio, P., Fiorito, G., Clayton, N. S., & Ostojić, L. (2019). Commentary: A conserved role for serotonergic neurotransmission in mediating social behavior in octopus. *Frontiers in Behavioral Neuroscience*, *13*. https://doi.org/10.3389/fnbeh.2019.00185

Anderson, J. R., & Gallup, G. G. (2011). Which primates recognize themselves in mirrors? *PLoS Biology*, *9*, e1001024. https://doi.org/10.1371/journal.pbio.1001024

Anderson, J. R., & Gallup, G. G. (2015). Mirror self-recognition: A review and critique of attempts to promote and engineer self-recognition in primates. *Primates*, *56*, 317–326. https://doi.org/10.1007/s10329-015-0488-9

Andrews, K. (2020). *How To Study Animal Minds*. Cambridge University Press https://doi.org/10.1017/9781108616522

Andrews, K., & Huss, B. (2014). Anthropomorphism, anthropectomy, and the null hypothesis. *Biology & Philosophy*, *29*, 711–729. https://doi.org/10.1007/s10539-014-9442-2

Anselme, P., & Robinson, M. J. F. (2019). Evidence for motivational enhancement of sign-tracking behavior under reward uncertainty. *Journal of Experimental Psychology: Animal Learning and Cognition*, *45*, 350–355. https://doi.org/10.1037/xan0000213

Aparecida Martins, R., Ribeiro Caldara, F., Crone, C., Markiy Odakura, A., Bevilacqua, A., Oliveira dos Santos Nieto, V. M., Aparecida Felix, G., Pereira dos Santos, A., Sousa dos Santos, L., Garófallo Garcia, R., & de Castro Lippi, I. C. (2021). Strategic use of straw as environmental enrichment for prepartum sows in farrowing crates. *Applied Animal Behaviour Science*, *234*, 105194. https://doi.org/10.1016/j.applanim.2020.105194

Aria, M., Alterisio, A., Scandurra, A., Pinelli, C., & D'Aniello, B. (2021). The scholar's best friend: Research trends in dog cognitive and behavioral studies. *Animal Cognition*, *24*, 541–553. https://doi.org/10.1007/s10071-020-01448-2

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

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

Balakhonov, D., & Rose, J. (2017). Crows rival monkeys in cognitive capacity. *Scientific Reports*, *7*, 8809. https://doi.org/10.1038/s41598-017-09400-0

Bandini, E., Bandini, M., & Tennie, C. (2021). A short report on the extent of stone handling behavior across otter species. *Animal Behavior and Cognition*, *8*, 15–22. https://doi.org/10.26451/abc.08.01.02.2021

Banks, G. C., O'Boyle, E. H., Pollack, J. M., White, C. D., Batchelor, J. H., Whelpley, C. E., Abston, K. A., Bennett, A. A., & Adkins, C. L. (2016). Questions about questionable research practices in the field of management: a guest commentary. *Journal of Management*, *42*, 5–20. https://doi.org/10.1177/0149206315619011

Bard, K. A., & Leavens, D. A. (2014). The importance of development for comparative primatology. Annual Review of Anthropology, 43, 183-200.

Baribault, B., Donkin, C., Little, D. R., Trueblood, J. S., Oravecz, Z., van Ravenzwaaij, D., White, C. N., De Boeck, P., & Vandekerckhove, J. (2018). Metastudies for robust tests of theory. *Proceedings of the National Academy of Sciences*, *115*, 2607–2612. https://doi.org/10.1073/pnas.1708285114

Barker, K. B., & Povinelli, D. J. (2019). Anthropomorphomania and the rise of the animal mind: a conversation. *Journal of Folklore Research*, *56*, 71–90. JSTOR.

Barrett, L. (2015). Why brains are not computers, why behaviorism is not satanism, and why dolphins are not aquatic apes. *The Behavior Analyst*, *39*, 9–23. https://doi.org/10.1007/s40614-015-0047-0

Bastos, A. P. M., Horváth, K., Webb, J. L., Wood, P. M., & Taylor, A. H. (2021). Self-care tooling innovation in a disabled kea (*Nestor notabilis*). *Scientific Reports*, *11*, 18035. https://doi.org/10.1038/s41598-021-97086-w

Bastos, A. P. M., & Taylor, A. H. (2020). Macphail's Null hypothesis of vertebrate intelligence: insights from avian cognition. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.01692

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

Bayne, T., Brainard, D., Byrne, R. W., Chittka, L., Clayton, N., Heyes, C., Mather, J., Ölveczky, B., Shadlen, M., Suddendorf, T., & Webb, B. (2019). What is cognition? *Current Biology*, *29*, R608–R615. https://doi.org/10.1016/j.cub.2019.05.044

BBSRC (2007) “BBSRC Data Sharing Policy: version 1.22 (March 2017 update)” URL: https://bbsrc.ukri.org/about/policies/policies-standards/data-sharing-policy/ Accessed on 29/09/2021

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

Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature*, *483*, 531–533. https://doi.org/10.1038/483531a

Bem, D. J. (2004). Writing the empirical journal article. In *The compleat academic: A career guide, 2nd ed* (pp. 185–219). American Psychological Association.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*, 407–425. https://doi.org/10.1037/a0021524

Beran, M. J. (2001). Summation and numerousness judgments of sequentially presented sets of items by chimpanzees (*Pan troglodytes*). *Journal of Comparative Psychology*, *115*, 181–191. https://doi.org/10.1037/0735-7036.115.2.181

Beran, M. J. (2006). Quantity perception by adult humans (*Homo sapiens*), chimpanzees (*Pan troglodytes*), and Rhesus macaques (*Macaca mulatta*) as a function of stimulus organization. *International Journal of Comparative Psychology*, *19*. https://escholarship.org/uc/item/76n9h06b

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

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

Beran, M. J. (2018). Replication and pre-registration in comparative psychology. *International Journal of Comparative Psychology*, Retrieved from https://escholarship.org/uc/item/59f4z2nd

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

Beran, M. J., French, K., Smith, T. R., & Parrish, A. E. (2019). Limited evidence of number–space mapping in rhesus monkeys (*Macaca mulatta*) and capuchin monkeys (*Sapajus apella*). *Journal of Comparative Psychology*, *133*, 281–293. https://doi.org/10.1037/com0000177

Beran, M. J., Parrish, A. E., Perdue, B. M., & Washburn, D. A. (2014). Comparative cognition: Past, present, and future. *International Journal of Comparative Psychology*, *27*, 3–30. Retrieved from https://escholarship.org/uc/item/9kh2m6rk

Billard, P., Schnell, A. K., Clayton, N. S., & Jozet-Alves, C. (2020). Cuttlefish show flexible and future-dependent foraging cognition. *Biology Letters*, *16*, 20190743. https://doi.org/10.1098/rsbl.2019.0743

Bird, C. D., & Emery, N. J. (2009). Rooks use stones to raise the water level to reach a floating worm. *Current Biology*, *19*, 1410–1414. https://doi.org/10.1016/j.cub.2009.07.033

Bishop, D. (2019). Rein in the four horsemen of irreproducibility. *Nature*, *568*, 435–435. https://doi.org/10.1038/d41586-019-01307-2

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

Blastland, M. (2019). *The hidden half: How the world conceals its secrets*. Atlantic Books.

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

Bloor, D. (1974). Popper's mystification of objective knowledge. *Science Studies*, *4*, 65–76. JSTOR.

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

Boesch, C. (2012). *The Ecology and Evolution of Social Behavior and Cognition in Primates*. The Oxford Handbook of Comparative Evolutionary Psychology. https://doi.org/10.1093/oxfordhb/9780199738182.013.0026

Boesch, C. (2021). Identifying animal complex cognition requires natural complexity. *IScience*, *24*, 102195. https://doi.org/10.1016/j.isci.2021.102195

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

Bohn, M., Eckert, J., Hanus, D., & Haun, D. B. M. (2021). *A longitudinal study of great ape cognition: stability, reliability and the influence of individual characteristics*. PsyArXiv. https://doi.org/10.31234/osf.io/pdt5w

Bolnick, D. (2021, May 12). Eco-Evo Evo-Eco: 17 months. *Eco-Evo Evo-Eco*. http://ecoevoevoeco.blogspot.com/2021/05/17-months.html

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*, 425. https://doi.org/10.1007/s11336-006-1447-6

Borsboom, D. (2014). *Theoretical Amnesia*. http://osc.centerforopenscience.org/2013/11/20/theoretical-amnesia/

Bourjade, M., Cochet, H., Molesti, S., & Guidetti, M. (2020). Is conceptual diversity an advantage for scientific inquiry? a case study on the concept of 'gesture' in comparative psychology. *Integrative Psychological and Behavioral Science*, *54*, 805–832. https://doi.org/10.1007/s12124-020-09516-5

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

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

Brecht, K. F., Müller, J., & Nieder, A. (2020). Carrion crows (*Corvus corone corone*) fail the mirror mark test yet again. *Journal of Comparative Psychology*, *134*, 372–378. https://doi.org/10.1037/com0000231

Brecht, K. F., Ostojić, L., Legg, E. W., & Clayton, N. S. (2018). Difficulties when using video playback to investigate social cognition in California scrub-jays (*Aphelocoma californica)*. *PeerJ*, *6*, e4451. https://doi.org/10.7717/peerj.4451

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

Brosnan, S. F., & de Waal, F. B. M. (2003). Monkeys reject unequal pay. *Nature*, *425*, 297–299. https://doi.org/10.1038/nature01963

Buckner, C. (2013). Morgan's Canon, meet Hume's Dictum: Avoiding anthropofabulation in cross-species comparisons. *Biology & Philosophy*, *28*, 853–871. https://doi.org/10.1007/s10539-013-9376-0

Bugnyar, T. (2011). Knower-guesser differentiation in ravens: Others' viewpoints matter. *Proceedings. Biological Sciences*, *278*, 634–640. https://doi.org/10.1098/rspb.2010.1514

Bugnyar, T., & Heinrich, B. (2006). Pilfering ravens, *Corvus corax*, adjust their behaviour to social context and identity of competitors. *Animal Cognition*, *9*, 369–376. https://doi.org/10.1007/s10071-006-0035-6

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

Bugnyar, T., Reber, S. A., & Buckner, C. (2016). Ravens attribute visual access to unseen competitors. *Nature Communications*, *7*, 10506. https://doi.org/10.1038/ncomms10506

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*, 511–517. https://doi.org/10.1111/j.1439-0310.2012.02040.x

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

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

Call, J., Burghardt, G. M., Pepperberg, I. M., Snowdon, C. T., & Zentall, T. (2017). What is comparative psychology? In *APA handbook of comparative psychology: Basic concepts, methods, neural substrate, and behavior, Vol. 1* (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*(9), 637. https://doi.org/10.1038/s41562-018-0399-z

Campbell, P. (2008). Escape from the impact factor. *Ethics in Science and Environmental Politics*, *8*, 5–7. https://doi.org/10.3354/esep00078

Canteloup, C., & Meunier, H. (2017). 'Unwilling' versus 'unable': Tonkean macaques' understanding of human goal-directed actions. *PeerJ*, *5*, e3227. https://doi.org/10.7717/peerj.3227

Carter, G. G., & Wilkinson, G. S. (2013). Food sharing in vampire bats: Reciprocal help predicts donations more than relatedness or harassment. *Proceedings. Biological Sciences*, *280*, 20122573. https://doi.org/10.1098/rspb.2012.2573

Cauchoix, M., Chow, P. K. Y., van Horik, J. O., Atance, C. M., Barbeau, E. J., Barragan-Jason, G., Bize, P., Boussard, A., Buechel, S. D., Cabirol, A., Cauchard, L., Claidière, N., Dalesman, S., Devaud, J. M., Didic, M., Doligez, B., Fagot, J., Fichtel, C., Henke-von der Malsburg, J., … Morand-Ferron, J. (2018). The repeatability of cognitive performance: A meta-analysis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*, 20170281. https://doi.org/10.1098/rstb.2017.0281

Chambers, C. (2017). *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice*. Princeton University Press.

Chambers, C. D. (2013). Registered Reports: A new publishing initiative at Cortex. *Cortex*, *49*, 609–610. https://doi.org/10.1016/j.cortex.2012.12.016

Chapelain, A. S., Hogervorst, E., Mbonzo, P., & Hopkins, W. D. (2011). Hand preferences for bimanual coordination in 77 bonobos (*Pan paniscus*): replication and extension. *International Journal of Primatology*, *32*, 491–510. https://doi.org/10.1007/s10764-010-9484-5

Chapman, C. A., Bicca-Marques, J. C., Calvignac-Spencer, S., Fan, P., Fashing, P. J., Gogarten, J., Guo, S., Hemingway, C. A., Leendertz, F., Li, B., Matsuda, I., Hou, R., Serio-Silva, J. C., & Chr. Stenseth, N. (2019). Games academics play and their consequences: How authorship, h-index and journal impact factors are shaping the future of academia. *Proceedings of the Royal Society B: Biological Sciences*, *286*, 20192047. https://doi.org/10.1098/rspb.2019.2047

Cheke, L. G., & Clayton, N. S. (2012). Eurasian jays (*Garrulus glandarius*) overcome their current desires to anticipate two distinct future needs and plan for them appropriately. *Biology Letters*, *8*, 171–175. https://doi.org/10.1098/rsbl.2011.0909

Chuard, P. J. C., Vrtílek, M., Head, M. L., & Jennions, M. D. (2019). Evidence that nonsignificant results are sometimes preferred: Reverse P-hacking or selective reporting? *PLOS Biology*, *17*, e3000127. https://doi.org/10.1371/journal.pbio.3000127

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

Cimarelli, G., Schoesswender, J., Vitiello, R., Huber, L., & Virányi, Z. (2021). Partial rewarding during clicker training does not improve naïve dogs' learning speed and induces a pessimistic-like affective state. *Animal Cognition*, *24*, 107–119. https://doi.org/10.1007/s10071-020-01425-9

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

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

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*, 1063–1072. https://doi.org/10.1007/s10071-019-01297-8

Clary, D., & Kelly, D. M. (2011). Cache protection strategies of a non-social food-caching corvid, Clark's nutcracker (*Nucifraga columbiana*). *Animal Cognition*, *14*, 735–744. https://doi.org/10.1007/s10071-011-0408-3

Clary, D., & Kelly, D. M. (2016). Graded mirror self-recognition by Clark's nutcrackers. *Scientific Reports*, *6*(1), 1–11. https://doi.org/10.1038/srep36459

Clary, D., Stow, M. K., Vernouillet, A., & Kelly, D. M. (2019). Mirror-mediated responses of California scrub jays (*Aphelocoma californica*) during a caching task and the mark test. *Ethology*, eth.12954. https://doi.org/10.1111/eth.12954

Clayton, N. (2012). Corvid cognition: Feathered apes. *Nature*, *484*, 453–454. https://doi.org/10.1038/484453a

Clayton, N. S., Dally, J. M., & Emery, N. J. (2007). Social cognition by food-caching corvids. The western scrub-jay as a natural psychologist. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *362*, 507–522. https://doi.org/10.1098/rstb.2006.1992

Clayton, N. S., & Dickinson, A. (1998). Episodic-like memory during cache recovery by scrub jays. *Nature*, *395*, 272–274. https://doi.org/10.1038/26216

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003. https://doi.org/10.1037/0003-066X.49.12.997

Colbourne, J. A. D., Auersperg, A. M. I., Lambert, M. L., Huber, L., & Völter, C. J. (2021). Extending the reach of tooling theory: a neurocognitive and phylogenetic perspective. *Topics in Cognitive Science*, tops.12554. https://doi.org/10.1111/tops.12554

Coles, N. A., Tiokhin, L., Scheel, A. M., Isager, P. M., & Lakens, D. (2018). The costs and benefits of replication studies. *Behavioral and Brain Sciences*, *41*, e124. https://doi.org/10.1017/S0140525X18000596

Commins, S. (2018). *Behavioural neuroscience*. Cambridge university press.

Cook, D. J., Mulrow, C. D., & Haynes, R. B. (1997). Systematic reviews: Synthesis of best evidence for clinical decisions. *Annals of Internal Medicine*, *126*, 376–380. https://doi.org/10.7326/0003-4819-126-5-199703010-00006

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

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

Craig, D. P. A., & Abramson, C. I. (2018). Ordinal pattern analysis in comparative psychology—A flexible alternative to null hypothesis significance testing using an observation oriented modeling paradigm. *International Journal of Comparative Psychology:* Retrieved from https://escholarship.org/uc/item/08w0c08s

Crandall, C. S., & Sherman, J. W. (2016). On the scientific superiority of conceptual replications for scientific progress. *Journal of Experimental Social Psychology*, *66*, 93–99. https://doi.org/10.1016/j.jesp.2015.10.002

Crosby, R. (2019). *A comparative investigation of the attribution of desires and preferences*. https://doi.org/10.17863/CAM.52387

Crosby, R., Legg, E., Brecht, K. F., Mendl, M., Ostojic, L., & Clayton, N. (2020). *Male Eurasian jays flexibly alter their food sharing in line with partners' choices*. https://doi.org/10.31234/osf.io/pr3ym

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

Culina, A., van den Berg, I., Evans, S., & Sánchez-Tójar, A. (2020). Low availability of code in ecology: A call for urgent action. *PLOS Biology*, *18*, e3000763. https://doi.org/10.1371/journal.pbio.3000763

Cumming, G. (2008). Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science*, *3*, 286–300. https://doi.org/10.1111/j.1745-6924.2008.00079.x

Cunningham, P. J., & Shahan, T. A. (2020). Delays to food-predictive stimuli do not affect suboptimal choice in rats. *Journal of Experimental Psychology: Animal Learning and Cognition*, *46*(4), 385–397. https://doi.org/10.1037/xan0000245

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

Dally, J. M., Clayton, N. S., & Emery, N. J. (2006). The behaviour and evolution of cache protection and pilferage. *Animal Behaviour*, *72*, 13–23. https://doi.org/10.1016/J.ANBEHAV.2005.08.020

Dally, J. M., Emery, N. J., & Clayton, N. S. (2006). Food-caching Western scrub-jays keep track of who was watching when. *Science (New York, N.Y.)*, *312*, 1662–1665. https://doi.org/10.1126/science.1126539

Danchev, V., Rzhetsky, A., & Evans, J. A. (2019). Centralized scientific communities are less likely to generate replicable results. *ELife*, *8*, e43094. https://doi.org/10.7554/eLife.43094

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

Davies, G. M., & Gray, A. (2015). Don't let spurious accusations of pseudoreplication limit our ability to learn from natural experiments (and other messy kinds of ecological monitoring). *Ecology and Evolution*, *5*, 5295–5304. https://doi.org/10.1002/ece3.1782

de Waal, F. B. M., & Ferrari, P. F. (2010). Towards a bottom-up perspective on animal and human cognition. *Trends in Cognitive Sciences*, *14*, 201–207. https://doi.org/10.1016/j.tics.2010.03.003

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

Despret, V., Buchanan, B., & Latour, B. (2016). *What would animals say if we asked the right questions?* University of Minnesota Press.

DeVries, M. S., Winters, C. P., & Jawor, J. M. (2020). Similarities in expression of territorial aggression in breeding pairs of northern cardinals, *Cardinalis cardinalis*. *Journal of Ethology*, *38*, 377–382. https://doi.org/10.1007/s10164-020-00659-x

Dickinson, A. (2012). Associative learning and animal cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*, 2733–2742. https://doi.org/10.1098/rstb.2012.0220

Dougherty, L. R., & Guillette, L. M. (2018). Linking personality and cognition: A meta-analysis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*. https://doi.org/10.1098/rstb.2017.0282

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., Nosek, B. A., & Johannesson, M. (2015). Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences*, *112*, 15343–15347. https://doi.org/10.1073/pnas.1516179112

Duhem, P. (1976). Physical Theory and Experiment. In S. G. Harding (Ed.), *Can Theories be Refuted?* (pp. 1–40). Springer Netherlands. https://doi.org/10.1007/978-94-010-1863-0_1

Dworkin, B. R., & Miller, N. E. (1986). Failure to replicate visceral learning in the acute curarized rat preparation. *Behavioral Neuroscience*, *100*, 299–314. https://doi.org/10.1037/0735-7044.100.3.299

Eaton, T., Hutton, R., Leete, J., Lieb, J., Robeson, A., & Vonk, J. (2018). Bottoms-up! Rejecting Top-down Human-centered Approaches in Comparative Psychology. *International Journal of Comparative Psychology*, *31*. Retrieved from: https://escholarship.org/uc/item/11t5q9wt

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., Baranski, E., Bernstein, M. J., Bonfiglio, D. B. V., Boucher, L., Brown, E. R., Budiman, N. I., Cairo, A. H., Capaldi, C. A., Chartier, C. R., Chung, J. M., Cicero, D. C., Coleman, J. A., Conway, J. G., … Nosek, B. A. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, *67*, 68–82. https://doi.org/10.1016/j.jesp.2015.10.012

Ebersole, C. R., Mathur, M. B., Baranski, E., Bart-Plange, D.-J., Buttrick, N. R., Chartier, C. R., Corker, K. S., Corley, M., Hartshorne, J. K., IJzerman, H., Lazarević, L. B., Rabagliati, H., Ropovik, I., Aczel, B., Aeschbach, L. F., Andrighetto, L., Arnal, J. D., Arrow, H., Babincak, P., … Nosek, B. A. (2020). Many Labs 5: testing pre-data-collection peer review as an intervention to increase replicability. *Advances in Methods and Practices in Psychological Science*, *3*(3), 309–331. https://doi.org/10.1177/2515245920958687

Eckert, J., Rakoczy, H., Duguid, S., Herrmann, E., & Call, J. (2021). The ape lottery: chimpanzees fail to consider spatial information when drawing statistical inferences. *Animal Behavior and Cognition*, *8*, 305–324. https://doi.org/10.26451/abc.08.03.01.2021

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*, 51–61. https://doi.org/10.1089/ees.2016.0223

Elliott, D. B. (2014). The impact factor: A useful indicator of journal quality or fatally flawed? *Ophthalmic and Physiological Optics*, *34*, 4–7. https://doi.org/10.1111/opo.12107

Emery, N. J. (2006). Cognitive ornithology: The evolution of avian intelligence. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *361*, 23–43. https://doi.org/10.1098/rstb.2005.1736

Emery, N. J., & Clayton, N. S. (2001). Effects of experience and social context on prospective caching strategies by scrub jays. *Nature*, *414*, 443–446. https://doi.org/10.1038/35106560

Emery, N. J., & Clayton, N. S. (2004). The mentality of crows: Convergent evolution of intelligence in corvids and apes. *Science*, *306*, 1903–1907. https://doi.org/10.1126/science.1098410

Emery, N. J., & Clayton, N. S. (2008). How to Build a Scrub-Jay that Reads Minds. In *Origins of the Social Mind* (pp. 65–97). Springer Japan. https://doi.org/10.1007/978-4-431-75179-3_4

Emery, N. J., Dally, J. M., & Clayton, N. S. (2004). Western scrub-jays (*Aphelocoma californica*) use cognitive strategies to protect their caches from thieving conspecifics. *Animal Cognition*, *7*(1), 37–43. https://doi.org/10.1007/s10071-003-0178-7

Emery N. J., Seed A. M., von Bayern A. M. P., & Clayton N.S., (2007). Cognitive adaptations of social bonding in birds. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1480), 489–505. https://doi.org/10.1098/rstb.2006.1991

Engber, D. (2017, June 7). *Daryl Bem Proved ESP Is Real. Which Means Science Is Broken.* Slate Magazine. https://slate.com/health-and-science/2017/06/daryl-bem-proved-esp-is-real-showed-science-is-broken.html

Eronen, M. I., & Bringmann, L. F. (2021). The theory crisis in psychology: how to move forward. *Perspectives on Psychological Science*, 174569162097058. https://doi.org/10.1177/1745691620970586

Etz, A., & Vandekerckhove, J. (2016). A Bayesian perspective on the reproducibility project: psychology. *PLOS ONE*, *11*, e0149794. https://doi.org/10.1371/journal.pone.0149794

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *PLOS ONE*, *5*, e10068. https://doi.org/10.1371/journal.pone.0010068

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*(3), 891–904. https://doi.org/10.1007/s11192-011-0494-7

Farmer, H. L., Murphy, G., & Newbolt, J. (2019). Change in stingray behaviour and social networks in response to the scheduling of husbandry events. *Journal of Zoo and Aquarium Research*, *7*(4), 203–209. https://doi.org/10.19227/jzar.v7i4.441

Farrah, K., Young, K., Tunis, M. C., & Zhao, L. (2019). Risk of bias tools in systematic reviews of health interventions: An analysis of PROSPERO-registered protocols. *Systematic Reviews*, *8*, 280. https://doi.org/10.1186/s13643-019-1172-8

Farrar, B. G. (2020). *Evidence of tool use in a seabird?* PsyArXiv. https://doi.org/10.31234/osf.io/463hk

Farrar, B. G., Altschul, D. M., Fischer, J., van der Mescht, J., Placi, S., Troisi, C. A., Vernouillet, A., Clayton, N. S., & Ostojic, L. (2020). Trialling meta-research in comparative cognition: claims and statistical inference in animal physical cognition. *Animal Behavior and Cognition*, *7*, 419–444. https://doi.org/10.26451/abc.07.03.09.2020

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

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

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*, 273–295. https://doi.org/10.26451/abc.08.02.14.2021

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

FAS Dean Smith Confirms Scientific Misconduct by Marc Hauser. (2010, August 20). Harvard Magazine. https://www.harvardmagazine.com/2010/08/harvard-dean-details-hauser-scientific-misconduct

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

Fidler, F., Burgman, M. A., Cumming, G., Buttrose, R., & Thomason, N. (2006). Impact of Criticism of Null-Hypothesis Significance Testing on Statistical Reporting Practices in Conservation Biology. *Conservation Biology*, *20*, 1539–1544.

Fiedler, K., Kutzner, F., & Krueger, J. I. (2012). the long way from α-error control to validity proper: problems with a short-sighted false-positive debate. *Perspectives on Psychological Science*, *7*, 661–669. https://doi.org/10.1177/1745691612462587

Fiedler, K., McCaughey, L., & Prager, J. (2021). Quo vadis, methodology? The key role of manipulation checks for validity control and quality of science. *Perspectives on Psychological Science*, *16*, 816–826. https://doi.org/10.1177/1745691620970602

Fiedler, K., & Prager, J. (2018). The regression trap and other pitfalls of replication science—Illustrated by the Report of the Open Science Collaboration. *Basic and Applied Social Psychology*, *40*, 115–124. https://doi.org/10.1080/01973533.2017.1421953

Fiedler, K., & Schwarz, N. (2016). Questionable research practices revisited. *Social Psychological and Personality Science*, *7*(1), 45–52. https://doi.org/10.1177/1948550615612150

Field, S. M., Hoekstra, R., Bringmann, L., & Ravenzwaaij, D. van. (2019). When and why to replicate: as easy as 1, 2, 3? *Collabra: Psychology*, *5*, 46. https://doi.org/10.1525/collabra.218

Fitzpatrick, S. (2008). Doing away with Morgan's Canon. *Mind & Language*, *23*, 224–246. https://doi.org/10.1111/j.1468-0017.2007.00338.x

Flake, J. K., & Fried, E. I. (2019). *Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them*. PsyArXiv. https://doi.org/10.31234/osf.io/hs7wm

Fong, E. A., & Wilhite, A. W. (2017). Authorship and citation manipulation in academic research. *PLOS ONE*, *12*, e0187394. https://doi.org/10.1371/journal.pone.0187394

Forrt. (2019). *Introducing a Framework for Open and Reproducible Research Training (FORRT)*. OSF Preprints https://doi.org/10.31219/osf.io/bnh7p

Forsell, E., Viganola, D., Pfeiffer, T., Almenberg, J., Wilson, B., Chen, Y., Nosek, B. A., Johannesson, M., & Dreber, A. (2018). Predicting replication outcomes in the Many Labs 2 study. *Journal of Economic Psychology*, S0167487018303283. https://doi.org/10.1016/j.joep.2018.10.009

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

Fragaszy, D. M., & Mangalam, M. (2018). Tooling. In *Advances in the Study of Behavior* (Vol. 50, pp. 177–241). Elsevier. https://doi.org/10.1016/bs.asb.2018.01.001

Francis, G. (2012). Too good to be true: Publication bias in two prominent studies from experimental psychology. *Psychonomic Bulletin & Review*, *19*, 151–156. https://doi.org/10.3758/s13423-012-0227-9

Fraser, D., Spooner, J. M., & Schuppli, C. A. (2017). "Everyday" knowledge and a new paradigm of animal research. *Animal Behavior and Cognition*, *4*, 502–505. https://doi.org/10.26451/abc.04.04.08.2017

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

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

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

Freeberg, T. M. (2020). On mark-test replication and mirror self-recognition in magpies. *Journal of Comparative Psychology*, *134*, 361–362. https://doi.org/10.1037/com0000256

Frias-Navarro, D., Pascual-Soler, M., Perezgonzalez, J., Monterde-i-Bort, H., & Pascual-Llobell, J. (2021). Spanish Scientists' Opinion about Science and Researcher Behavior. *The Spanish Journal of Psychology*, *24*, e7. https://doi.org/10.1017/SJP.2020.59

Frolov, S. (2021). Quantum computing's reproducibility crisis: Majorana fermions. *Nature*, *592*, 350–352. https://doi.org/10.1038/d41586-021-00954-8

Geach, P. T. (1973). Ontological Relativity and Relative Identity. In M. K. Munitz (Ed.), *Logic and Ontology*. New York: New York University Press.

Gelman, A. (2014, September 5). *Confirmationist and falsificationist paradigms of science*. Statistical Modeling, Causal Inference, and Social Science. https://statmodeling.stat.columbia.edu/2014/09/05/confirmationist-falsificationist-paradigms-science/

Gelman, A. (2018). Don't characterize replications as successes or failures. *Behavioral and Brain Sciences*, *41*. https://doi.org/10.1017/S0140525X18000638

Gelman, A., & Carlin, J. (2014). Beyond power calculations. *Perspectives on Psychological Science*, *9*, 641–651. https://doi.org/10.1177/1745691614551642

Gelman, A., & Geurts, H. M. (2017). The statistical crisis in science: How is it relevant to clinical neuropsychology? *The Clinical Neuropsychologist*, *31*, 1000–1014. https://doi.org/10.1080/13854046.2016.1277557

Ghirlanda, S. (2017). Can squirrel monkeys learn an AB^nA grammar? A re-evaluation of Ravignani *et al.* (2013). *PeerJ*, *5*, e3806. https://doi.org/10.7717/peerj.3806

Ghirlanda, S., Lind, J., & Enquist, M. (2017). Memory for stimulus sequences: A divide between humans and other animals? *Royal Society Open Science*, *4*, 161011. https://doi.org/10.1098/rsos.161011

Gibbs, N. M., & Gibbs, S. V. (2015). Misuse of 'trend' to describe 'almost significant' differences in anaesthesia research. *BJA: British Journal of Anaesthesia*, *115*, 337–339. https://doi.org/10.1093/bja/aev149

Gigerenzer, G. (1998a). Surrogates for theories. *Theory & Psychology*, *8*, 195–204. https://doi.org/10.1177/0959354398082006

Gigerenzer, G. (1998b). We need statistical thinking, not statistical rituals. *Behavioral and Brain Sciences*, *21*(2), 199–200. https://doi.org/10.1017/S0140525X98281167

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*(5), 587–606. https://doi.org/10.1016/j.socec.2004.09.033

Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The null ritual: What you always wanted to know about null hypothesis testing but were afraid to ask. *Handbook on Quantitative Methods in the Social Sciences. Sage, Thousand Oaks, CA*, 389–406.

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on "Estimating the reproducibility of psychological science." *Science*, *351*, 1037–1037. https://doi.org/10.1126/science.aad7243

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

Goodman, S. (2008). A dirty dozen: twelve p-value misconceptions. *Seminars in Hematology*, *45*(3), 135–140. https://doi.org/10.1053/j.seminhematol.2008.04.003

Goodwin, D. (1956). Further Observations on the Behaviour of the jay *Garrulus glandarius*. *Ibis*, *98*(2), 186–219. https://doi.org/10.1111/j.1474-919X.1956.tb03040.x

Guadarrama, S. S., Domínguez-Vega, H., Díaz-Albiter, H. M., Quijano, A., Bastiaans, E., Carrillo-Castilla, P., Manjarrez, J., Gómez-Ortíz, Y., & Fajardo, V. (2020). Hypoxia by altitude and welfare of captive beaded lizards (*Heloderma Horridum*) in Mexico: hematological approaches. *Journal of Applied Animal Welfare Science*, *23*, 74–82. https://doi.org/10.1080/10888705.2018.1562350

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

Guyatt, G. H., Sackett, D. L., Sinclair, J. C., Hayward, R., Cook, D. J., Cook, R. J., Bass, E., Gerstein, H., Haynes, B., Holbrook, A., Jaeschke, R., Laupacls, A., Moyer, V., & Wilson, M. (1995). Users' guides to the medical literature: ix. a method for grading health care recommendations. *JAMA*, *274*, 1800–1804. https://doi.org/10.1001/jama.1995.03530220066035

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

Halperin, I., Vigotsky, A., Foster, C., & Pyne, D. (2018). Strengthening the practice of exercise and sport-science research. *International Journal of Sports Physiology and Performance*. https://doi.org/10.1123/ijspp.2017-0322

Hampton, R. (2019). Parallel overinterpretation of behavior of apes and corvids. *Learning & Behavior*, *47*, 105–106. https://doi.org/10.3758/s13420-018-0330-5

Hare, B., Plyusnina, I., Ignacio, N., Schepina, O., Stepika, A., Wrangham, R., & Trut, L. (2005). Social cognitive evolution in captive foxes is a correlated by-product of experimental domestication. *Current Biology*, *15*, 226–230. https://doi.org/10.1016/j.cub.2005.01.040

Harris, J. A., & Bouton, M. E. (2020). Pavlovian conditioning under partial reinforcement: The effects of nonreinforced trials versus cumulative conditioned stimulus duration. *Journal of Experimental Psychology: Animal Learning and Cognition*, *46*, 256–272. https://doi.org/10.1037/xan0000242

Hartgerink, C. H. J. (2016). 688,112 statistical results: Content mining psychology articles for statistical test results. *Data Archiving and Networked Services (DANS).* https://doi.org/10.17026/DANS-2CM-V9J9

Hashmi, A., & Sullivan, M. (2020). The visitor effect in zoo-housed apes: The variable effect on behaviour of visitor number and noise. *Journal of Zoo and Aquarium Research*, *8*(4), 268–282. https://doi.org/10.19227/jzar.v8i4.523

Hauser, M. D., Weiss, D., & Marcus, G. (2002). RETRACTED: Rule learning by cotton-top tamarins. *Cognition*, *86*(1), B15–B22. https://doi.org/10.1016/S0010-0277(02)00139-7

Haven, T. L., Bouter, L. M., Smulders, Y. M., & Tijdink, J. K. (2019). Perceived publication pressure in Amsterdam: Survey of all disciplinary fields and academic ranks. *PLOS ONE*, *14*, e0217931. https://doi.org/10.1371/journal.pone.0217931

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of *p*-hacking in science. *PLoS Biology*, *13*(3), e1002106. https://doi.org/10.1371/journal.pbio.1002106

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*, 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

Hedges, L. V. (1984). Estimation of effect size under nonrandom sampling: the effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics*, *9*(1), 61. https://doi.org/10.2307/1164832

Hemmer, B. M., Parrish, A. E., Wise, T. B., Davis, M., Branham, M., Martin, D. E., & Templer, V. L. (2019). Social vs. nonsocial housing differentially affects perseverative behavior in rats (*Ratus norvegicus*). *Animal Behavior and Cognition*, *6*(3), 168–178. https://doi.org/10.26451/abc.06.03.02.2019

Hempel, C. G. (1958). *The theoretician's dilemma: A study in the logic of theory construction*. http://conservancy.umn.edu/handle/11299/184621

Hennefield, L., Hwang, H. G., & Povinelli, D. J. (2019). Going meta: retelling the scientific retelling of Aesop's the crow and the pitcher. *Journal of Folklore Research*, *56*, 45–70. JSTOR.

Hennefield, L., Hwang, H. G., Weston, S. J., & Povinelli, D. J. (2018). Meta-analytic techniques reveal that corvid causal reasoning in the Aesop's Fable paradigm is driven by trial-and-error learning. *Animal Cognition*, *21*, 1–14. https://doi.org/10.1007/s10071-018-1206-y

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, *466*(7302), 29–29. https://doi.org/10.1038/466029a

Herrera-Bennett, A. C., Heene, M., Lakens, D., & Ufer, S. (2020). *Improving statistical inferences: Can a MOOC reduce statistical misconceptions about p-values, confidence intervals, and Bayes factors?* PsyArXiv. https://doi.org/10.31234/osf.io/zt3g9

Herrmann, E., Hare, B., Call, J., & Tomasello, M. (2010). Differences in the cognitive skills of bonobos and chimpanzees. *PLoS ONE*, *5*, e12438. https://doi.org/10.1371/journal.pone.0012438

Heyes, C. (2012). What's social about social learning? *Journal of Comparative Psychology*, *126*, 193–202. https://doi.org/10.1037/a0025180

Heyes, C. (2015). Animal mindreading: What's the problem? *Psychonomic Bulletin & Review*, *22*, 313–327. https://doi.org/10.3758/s13423-014-0704-4

Heyes, C. (2017). Apes submentalise. *Trends in Cognitive Sciences*, *21*, 1–2. https://doi.org/10.1016/j.tics.2016.11.006

Heyes, C. (2019). What is cognition? *Current Biology*, *29*, R608–R615. https://doi.org/10.1016/j.cub.2019.05.044

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

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

Hill, H. M. (2017). The Psychology of Cows? A case of over-interpretation and personification. *Animal Behavior and Cognition*, *4*, 506–511. https://doi.org/10.26451/abc.04.04.09.2017

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

Hoekstra, R., Finch, S., Kiers, H. A. L., & Johnson, A. (2006). Probability as certainty: Dichotomous thinking and the misuse of p values. *Psychonomic Bulletin & Review*, *13*, 1033–1037. https://doi.org/10.3758/BF03213921

Holekamp, K. E., Sakai, S. T., & Lundrigan, B. L. (2007). Social intelligence in the spotted hyena (*Crocuta crocuta*). *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*, 523–538. https://doi.org/10.1098/rstb.2006.1993

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*, e1002190. https://doi.org/10.1371/journal.pbio.1002190

Hopkins, W. D., Wesley, M. J., Izard, M. K., Hook, M., & Schapiro, S. J. (2004). Chimpanzees (*Pan troglodytes*) are predominantly right-handed: replication in three populations of apes. *Behavioral Neuroscience*, *118*(3), 659–663. https://doi.org/10.1037/0735-7044.118.3.659

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

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

Inoue, S., & Matsuzawa, T. (2007). Working memory of numerals in chimpanzees. *Current Biology*, *17*, R1004–R1005. https://doi.org/10.1016/j.cub.2007.10.027

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

Ioannidis, J. P. A. (2012a). Why science is not necessarily self-correcting. *Perspectives on Psychological Science*, *7*, 645–654. JSTOR.

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

Ioannidis, J. P. A. (2018). Meta-research: Why research on research matters. *PLOS Biology*, *16*, e2005468. https://doi.org/10.1371/journal.pbio.2005468

Ioannidis, J. P. A., Fanelli, D., Dunne, D. D., & Goodman, S. N. (2015). Meta-research: evaluation and improvement of research methods and practices. *PLOS Biology*, *13*, e1002264. https://doi.org/10.1371/journal.pbio.1002264

Isager, P. M., Veer, A. van ’t, & Lakens, D. (2021). *Replication value as a function of citation impact and sample size*. MetaArXiv. https://doi.org/10.31222/osf.io/knjea

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

Jones, J. O., Moody, W. M., & Shields, J. D. (2021). Microenvironmental modulation of the developing tumour: an immune-stromal dialogue. Molecular Oncology, 15(10), 2600-2633.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. Psychological Science, 23, 524–532. https://doi.org/10.1177/0956797611430953Johnson, Z., Brent, L., Alvarenga, J. C., Comuzzie, A. G., Shelledy, W., Ramirez, S., Cox, L., Mahaney, M. C., Huang, Y.-Y., Mann, J. J., Kaplan, J. R., & Rogers, J. (2015). Genetic influences on response to novel objects and dimensions of personality in Papio baboons. *Behavior Genetics*, *45*, 215–227. https://doi.org/10.1007/s10519-014-9702-6

Jonsson, M., Ghirlanda, S., Lind, J., Vinken, V., & Enquist, M. (2021). Learning Simulator: A simulation software for animal and human learning. *Journal of Open Source Software*, *6*(58), 2891. https://doi.org/10.21105/joss.02891

Jozet-Alves, C., Bertin, M., & Clayton, N. S. (2013). Evidence of episodic-like memory in cuttlefish. *Current Biology*, *23*, R1033–R1035. https://doi.org/10.1016/j.cub.2013.10.021

Jussim, L., Stevens, S. T., Honeycutt, N., Anglin, S. M., Fox, N., Stevens, S. T., Honeycutt, N., Anglin, S. M., & Fox, N. (2019). *Scientific gullibility*. The Social Psychology of Gullibility; Routledge. https://doi.org/10.4324/9780429203787-15

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

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

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

Kafkafi, N., Benjamini, Y., Sakov, A., Elmer, G. I., & Golani, I. (2005). Genotype-environment interactions in mouse behavior: A way out of the problem. *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 4619–4624. https://doi.org/10.1073/pnas.0409554102

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

Kaminski, J., Call, J., & Fischer, J. (2004). Word learning in a domestic dog: evidence for "fast mapping." *Science*, *304*, 1682–1683. https://doi.org/10.1126/science.1097859

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

Karp, N. A., Speak, A. O., White, J. K., Adams, D. J., Hrabé de Angelis, M., Hérault, Y., & Mott, R. F. (2014). Impact of temporal variation on design and analysis of mouse knockout phenotyping studies. *PloS One*, *9*, e111239. https://doi.org/10.1371/journal.pone.0111239

Kawaguchi, M., & Kuriwada, T. (2020). Effect of predator cue on escape and oviposition behaviour of freshwater snail. *Behaviour*, *157*, 683–697. https://doi.org/10.1163/1568539X-bja10018

Kawai, N., Nakagami, A., Yasue, M., Koda, H., & Ichinohe, N. (2019). Common marmosets (*Callithrix jacchus*) evaluate third-party social interactions of human actors but Japanese monkeys (*Macaca fuscata*) do not. *Journal of Comparative Psychology*, *133*, 488–495. https://doi.org/10.1037/com0000182

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology, Inc*, *2*, 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Khan, N., & Wascher, C. A. F. (2021). Considering generalizability: a lesson from auditory enrichment research on zoo animals. *Animal Behavior and Cognition*, *8*, 251–262. https://doi.org/10.26451/abc.08.02.12.2021

Kirchhofer, K. C., Zimmermann, F., Kaminski, J., & Tomasello, M. (2012). Dogs (*Canis familiaris*), but not chimpanzees (*Pan troglodytes*), understand imperative pointing. *PLoS ONE*, *7*, e30913. https://doi.org/10.1371/journal.pone.0030913

Kirschhock, M. E., Ditz, H. M., & Nieder, A. (2021). Behavioral and neuronal representation of numerosity zero in the crow. *Journal of Neuroscience*, *41*, 4889–4896. https://doi.org/10.1523/JNEUROSCI.0090-21.2021

Klayman, J., & Ha, Y. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review*, *94*, 211–228. https://doi.org/10.1037/0033-295X.94.2.211

Klein, R. A., Cook, C. L., Ebersole, C. R., Vitiello, C. A., Nosek, B. A., Chartier, C. R., Christopherson, C. D., Clay, S., Collisson, B., Crawford, J., Cromar, R., Vidamuerte, D., Gardiner, G., Gosnell, C., Grahe, J. E., Hall, C., Joy-Gaba, J. A., Legg, A. M., Levitan, C., … Ratliff, K. A. (2019). *Many Labs 4: failure to replicate mortality salience effect with and without original author involvement*. PsyArXiv. https://doi.org/10.31234/osf.io/vef2c

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., … Nosek, B. A. (2014). Investigating variation in replicability: A "Many Labs" replication project. *Social Psychology*, *45*, 142–152. https://doi.org/10.1027/1864-9335/a000178

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., … Nosek, B. A. (2018). Many labs 2:

investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*, 443–490. https://doi.org/10.1177/2515245918810225

Koczura, M., Martin, B., Musci, M., Massimo, M. D., Bouchon, M., Turille, G., Kreuzer, M., Berard, J., & Coppa, M. (2021). Little difference in milk fatty acid and terpene composition among three contrasting dairy breeds when grazing a biodiverse mountain pasture. *Frontiers in Veterinary Science*, *7*, 612504. https://doi.org/10.3389/fvets.2020.612504

Koroshetz, W. J., Behrman, S., Brame, C. J., Branchaw, J. L., Brown, E. N., Clark, E. A., Dockterman, D., Elm, J. J., Gay, P. L., Green, K. M., Hsi, S., Kaplitt, M. G., Kolber, B. J., Kolodkin, A. L., Lipscombe, D., MacLeod, M. R., McKinney, C. C., Munafò, M. R., Oakley, B., … Silberberg, S. D. (2020). Framework for advancing rigorous research. *ELife*, *9*, e55915. https://doi.org/10.7554/eLife.55915

Kousta, S. (2021). The value of evidence synthesis. *Nature Human Behaviour*, *5*, 539–539. https://doi.org/10.1038/s41562-021-01131-7

Krämer, U. (2015). Planting molecular functions in an ecological context with Arabidopsis thaliana. *ELife*, *4*, e06100. https://doi.org/10.7554/eLife.06100

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

Krebs, J. R., & Davies, N. B. (Eds.). (1997). *Behavioural ecology: An evolutionary approach* (4th ed). Blackwell Science.

Krupenye, C., Kano, F., Hirata, S., Call, J., & Tomasello, M. (2016). Great apes anticipate that other individuals will act according to false beliefs. *Science*, *354*, 110–114. https://doi.org/10.1126/science.aaf8110

Krupenye, C., Kano, F., Hirata, S., Call, J., & Tomasello, M. (2017). A test of the submentalizing hypothesis: Apes' performance in a false belief task inanimate control. *Communicative & Integrative Biology*, *10*. https://doi.org/10.1080/19420889.2017.1343771

Kvarnemo, C., Andersson, S. E., Elisson, J., Moore, G. I., & Jones, A. G. (2021). Home range use in the West Australian seahorse *Hippocampus subelongatus* is influenced by sex and partner's home range but not by body size or paired status. *Journal of Ethology*, *39*, 235–248. https://doi.org/10.1007/s10164-021-00698-y

Lakatos, I. (1970). Falsification and the Methodology of Scientific Research Programmes. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the Growth of Knowledge* (pp. 91–196). Cambridge University Press. https://doi.org/10.1017/CBO9781139171434.009

Lakens, D. (2015). On the challenges of drawing conclusions from *p* -values just below 0.05. *PeerJ*, *3*, e1142. https://doi.org/10.7717/peerj.1142

Lakens, D. (2017). Equivalence tests: a practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *8*, 355–362. https://doi.org/10.1177/1948550617697177

Lakens, D. (2021). *Sample Size Justification* . PsyArXiv. https://doi.org/10.31234/osf.io/9d3yf

Lambdin, C. (2012). Significance tests as sorcery: Science is empirical—significance tests are not. *Theory & Psychology*, *22*, 67–90. https://doi.org/10.1177/0959354311429854

Lambert, M., Farrar, B. G., Garcia-Pelegrin, E., Reber, S. A., & Miller, R. (2021). *ManyBirds: A multi-site collaborative Open Science approach to avian cognition and behaviour research*. PsyArXiv. https://doi.org/10.31234/osf.io/83xkt

Lane, D. M., & Dunlap, W. P. (1978). Estimating effect size: Bias resulting from the significance criterion in editorial decisions. *British Journal of Mathematical and Statistical Psychology*, *31*, 107–112. https://doi.org/10.1111/j.2044-8317.1978.tb00578.x

Lane, J. (2010). Let's make science metrics more scientific. *Nature*, *464*, 488–489. https://doi.org/10.1038/464488a

Lange, F. (2019). Are difficult-to-study populations too difficult to study in a reliable way? *European Psychologist*, 1–10. https://doi.org/10.1027/1016-9040/a000384

LastWeekTonight. (2016, May 9). *Scientific Studies: Last Week Tonight with John Oliver (HBO)*. https://www.youtube.com/watch?v=0Rnq1NpHdmw

Latour, B. (1987). *Science in action: How to follow scientists and engineers through society*. Harvard University Press.

Latour, B., & Woolgar, S. (1986). *Laboratory life: The construction of scientific facts*. Princeton University Press.

Lawrence, J. M., Meyerowitz-Katz, G., Heathers, J. A. J., Brown, N. J. L., & Sheldrick, K. A. (2021). The lesson of ivermectin: Meta-analyses based on summary data alone are inherently unreliable. *Nature Medicine*, 1–2. https://doi.org/10.1038/s41591-021-01535-y

Lazarowski, L., Thompkins, A., Krichbaum, S., Waggoner, L. P., Deshpande, G., & Katz, J. S. (2020). Comparing pet and detection dogs (*Canis familiaris*) on two aspects of social cognition. *Learning & Behavior*, *48*, 432–443. https://doi.org/10.3758/s13420-020-00431-8

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

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

Lazic, S. E. (2016). *Experimental Design for Laboratory Biologists: Maximising Information and Improving Reproducibility* (1st ed.). Cambridge University Press. https://doi.org/10.1017/9781139696647

Lazic, S. E. (2021). Why multiple hypothesis test corrections provide poor control of false positives in the real world. *ArXiv:2108.04752*. http://arxiv.org/abs/2108.04752

Lazic, S. E., Clarke-Williams, C. J., & Munafò, M. R. (2018). What exactly is 'N' in cell culture and animal experiments? *PLOS Biology*, *16*, e2005282. https://doi.org/10.1371/journal.pbio.2005282

Leavens, D. A., Bard, K. A., & Hopkins, W. D. (2010). BIZARRE chimpanzees do not represent "the chimpanzee". Behavioral and Brain Sciences, 33(2-3), 100-101.

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

Lefebvre, L., Reader, S. M., & Sol, D. (2004). Brains, Innovations and evolution in birds and primates. *Brain, Behavior and Evolution*, *63*, 233–246. https://doi.org/10.1159/000076784

Legg, E. W., & Clayton, N. S. (2014). Eurasian jays (*Garrulus glandarius*) conceal caches from onlookers. *Animal Cognition*, *17*, 1223–1226. https://doi.org/10.1007/s10071-014-0743-2

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

Lewis, D. K. (1993). Many, but Almost One. In K. Cambell, J. Bacon, & L. Reinhardt (Eds.), *Ontology, Causality and Mind: Essays on the Philosophy of D. M. Armstrong* (pp. 23–38). Cambridge University Press.

Lilienfeld, S. O. (2017). Psychology's replication crisis and the grant culture: righting the ship. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *12*, 660–664. https://doi.org/10.1177/1745691616687745

Lilley, M. K., de Vere, A. J., & Yeater, D. B. (2020). Laterality of eye use by bottlenose (*Tursiops truncatus*) and rough-toothed (*Steno bredanensis*) dolphins while viewing predictable and unpredictable stimuli. *International Journal of Comparative Psychology*, *33*. https://doi.org/10.46867/ijcp.2020.33.03.01

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

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

Lit, L., Schweitzer, J. B., & Oberbauer, A. M. (2011). Handler beliefs affect scent detection dog outcomes. *Animal Cognition*, *14*, 387–394. https://doi.org/10.1007/s10071-010-0373-2

Little, D. R., & Smith, P. L. (2018). Replication is already mainstream: Lessons from small-N designs. *Behavioral and Brain Sciences*, *41*. https://doi.org/10.1017/S0140525X18000766

Llorente, M., Riba, D., Palou, L., Carrasco, L., Mosquera, M., Colell, M., & Feliu, O. (2011). Population-level right-handedness for a coordinated bimanual task in naturalistic housed chimpanzees: Replication and extension in 114 animals from Zambia and Spain. *American Journal of Primatology*, *73*, 281–290. https://doi.org/10.1002/ajp.20895

Locey, M. L. (2020). The evolution of behavior analysis: toward a replication crisis? *Perspectives on Behavior Science*, *43*, 655–675. https://doi.org/10.1007/s40614-020-00264-w

Loehle, C. (1987). Hypothesis testing in ecology: psychological aspects and the importance of theory maturation. *The Quarterly Review of Biology*, *62*, 397–409. https://doi.org/10.1086/415619

Maatman, F. O. (2021). *Psychology's Theory Crisis, and Why Formal Modelling Cannot Solve It*. PsyArXiv. https://doi.org/10.31234/osf.io/puqvs

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

MacLean, E. L., Hare, B., Nunn, C. L., Addessi, E., Amici, F., Anderson, R. C., Aureli, F., Baker, J. M., Bania, A. E., Barnard, A. M., Boogert, N. J., Brannon, E. M., Bray, E. E., Bray, J., Brent, L. J. N., Burkart, J. M., Call, J., Cantlon, J. F., Cheke, L. G., … Zhao, Y. (2014). The evolution of self-control. *Proceedings of the National Academy of Sciences of the United States of America*, *111*, E2140-2148. https://doi.org/10.1073/pnas.1323533111

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

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

Maes, E., Boddez, Y., Alfei, J. M., Krypotos, A.-M., D'Hooge, R., De Houwer, J., & Beckers, T. (2016). The elusive nature of the blocking effect: 15 failures to replicate. *Journal of Experimental Psychology. General*, *145*, e49-71. https://doi.org/10.1037/xge0000200

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

Many Primates, Altschul, D., Beran, M. J., Bohn, M., Caspar, K., Fichtel, C., Försterling, M., Grebe, N., Hernandez-Aguilar, R. A., Kwok, S. C., Rodrigo, A. M., Proctor, D., Sanchez-Amaro, A., Simpson, E. A., Szabelska, A., Taylor, D., van der Mescht, J., Völter, C., & Watzek, J. (2019). Collaborative open science as a way to reproducibility and new insights in primate cognition research *Japanese Psychological Review* 62:205220

Many Primates, Altschul, D. M., Beran, M. J., Bohn, M., Call, J., DeTroy, S., Duguid, S. J., Egelkamp, C. L., Fichtel, C., Fischer, J., Flessert, M., Hanus, D., Haun, D. B. M., Haux, L. M., Hernandez-Aguilar, R. A., Herrmann, E., Hopper, L. M., Joly, M., Kano, F., … Watzek, J. (2019). Establishing an infrastructure for collaboration in primate cognition research. *PLOS ONE*, *14*, e0223675. https://doi.org/10.1371/journal.pone.0223675

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

Massen, J. J. M., Ritter, C., & Bugnyar, T. (2015). Tolerance and reward equity predict cooperation in ravens (*Corvus corax*). *Scientific Reports*, *5*, 15021. https://doi.org/10.1038/srep15021

Mayo, D. G. (2018). *Statistical inference as severe testing: How to get beyond the statistics wars*. Cambridge University Press.

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

McElreath, R., & Smaldino, P. E. (2015). Replication, communication, and the population dynamics of scientific discovery. *PLOS ONE*, *10*, e0136088. https://doi.org/10.1371/journal.pone.0136088

McShane, B. B., & Gal, D. (2016). Blinding us to the obvious? the effect of statistical training on the evaluation of evidence. *Management Science*, *62*, 1707–1718. https://doi.org/10.1287/mnsc.2015.2212

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon statistical significance. *The American Statistician*, *73*, 235–245. https://doi.org/10.1080/00031305.2018.1527253

Meehl, P. E. (1967). Theory-testing in psychology and physics: a methodological paradox. *Philosophy of Science*, *34*, 103–115. JSTOR.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*, 806–834. https://doi.org/10.1037/0022-006X.46.4.806

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, *66*(1), 195–244. https://doi.org/10.2466/pr0.1990.66.1.195

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

Meza, P., Elias, D. O., & Rosenthal, M. F. (2021). The effect of substrate on prey capture does not match natural substrate use in a wolf spider. *Animal Behaviour*, *176*, 17–21. https://doi.org/10.1016/j.anbehav.2021.03.014

Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, *88*, 355–383. https://doi.org/10.1111/j.2044-8295.1997.tb02641.x

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

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

Mitchell, J. P. (2014). On the evidentiary emptiness of failed replications. *Http://Jasonmitchell.Fas.Harvard.Edu/Papers/Mitchell_failed_science_2014.Pdf*.

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

Morey, R., & Lakens, D. (2016). *Why most of psychology is statistically unfalsifiable*. https://doi.org/10.5281/zenodo.838685

Motes Rodrigo, A., Ramirez Torres, C. E., Hernandez Salazar, L. T., & Laska, M. (2018). Hand preferences in two unimanual and two bimanual coordinated tasks in the black-handed

spider monkey (*Ateles geoffroyi*). *Journal of Comparative Psychology*, *132*, 220–229. https://doi.org/10.1037/com0000110

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

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

Murphey, R. M. (1967). Instrumental conditioning of the fruit fly, *Drosophila melanogaster*. *Animal Behaviour*, *15*, 153–161. https://doi.org/10.1016/s0003-3472(67)80027-7

National Institute for Health and Care Excellence (2012), Appendix C: Methodology checklist: randomised controlled trials [The Guidelines Manual] https://www.nice.org.uk/process/pmg6/resources/the-guidelines-manual-appendices-bi-2549703709/chapter/appendix-c-methodology-checklist-randomised-controlled-trials

Nelson, E. L., Figueroa, A., Albright, S. N., & Gonzalez, M. F. (2015). Evaluating handedness measures in spider monkeys. *Animal Cognition*, *18*, 345–353. https://doi.org/10.1007/s10071-014-0805-5

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

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

Neyman, J. (1976). Tests of statistical hypotheses and their use in studies of natural phenomena. *Communications in Statistics - Theory and Methods*, *5*, 737–751. https://doi.org/10.1080/03610927608827392

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *231*, 289–337. https://doi.org/10.1098/rsta.1933.0009

Nickerson, R. S. (1998). Confirmation Bias: a ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*, 175–220. https://doi.org/10.1037/1089-2680.2.2.175

Nieder, A., Wagener, L., & Rinnert, P. (2020). A neural correlate of sensory consciousness in a corvid bird. *Science*, *369*, 1626–1629. https://doi.org/10.1126/science.abb1447

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

Nissen, S. B., Magidson, T., Gross, K., & Bergstrom, C. T. (2016). Publication bias and the canonization of false facts. *ELife*, *5*, e21451. https://doi.org/10.7554/eLife.21451

Noonan, H., & Curtis, B. (2004). *Identity*. https://plato.stanford.edu/archives/sum2018/entries/identity/

Nord, C. L., Valton, V., Wood, J., & Roiser, J. P. (2017). Power-up: a reanalysis of "power failure" in neuroscience using mixture modeling. *Journal of Neuroscience*, *37*, 8051–8061. https://doi.org/10.1523/JNEUROSCI.3592-16.2017

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

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline Struhl, M., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2021). *Replicability, Robustness, and Reproducibility in Psychological Science*. PsyArXiv. https://doi.org/10.31234/osf.io/ksfvq

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology*, *45*, 137–141. https://doi.org/10.1027/1864-9335/a000192

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

Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, *48*, 1205–1226. https://doi.org/10.3758/s13428-015-0664-2

Nuijten, M. B., & Polanin, J. R. (2020). "statcheck": Automatically detect statistical reporting inconsistencies to increase reproducibility of meta-analyses. *Research Synthesis Methods*, *11*, 574–579. https://doi.org/10.1002/jrsm.1408

Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-019-01645-2

O'Connor, C., & Weatherall, J. O. (2020). False beliefs and the social structure of science: some models and case studies. In D. M. Allen & J. W. Howell (Eds.), *Groupthink in Science: Greed, Pathological Altruism, Ideology, Competition, and Culture* (pp. 37–48). Springer International Publishing. https://doi.org/10.1007/978-3-030-36822-7_4

O'Donoghue, E. M., Broschard, M. B., & Wasserman, E. A. (2020). Pigeons exhibit flexibility but not rule formation in dimensional learning, stimulus generalization, and task switching. *Journal of Experimental Psychology: Animal Learning and Cognition*, *46*, 107–123. https://doi.org/10.1037/xan0000234

Olkowicz, S., Kocourek, M., Lučan, R. K., Porteš, M., Fitch, W. T., Herculano-Houzel, S., & Němec, P. (2016). Birds have primate-like numbers of neurons in the forebrain. *Proceedings of the National Academy of Sciences*, *113*, 7255–7260. https://doi.org/10.1073/pnas.1517131113

Olmstead, M. C., & Kuhlmeier, V. A. (2015). *Comparative Cognition*. Cambridge University Press. https://doi.org/10.1017/CBO9780511894787

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

Oostenbroek, J., Suddendorf, T., Nielsen, M., Redshaw, J., Kennedy-Costantini, S., Davis, J., ... & Slaughter, V. (2016). Comprehensive longitudinal study challenges the existence of neonatal imitation in humans. Current biology, 26(10), 1334-1338.

Open Science Collaboration, O. S. (2015). PSYCHOLOGY. Estimating the reproducibility of psychological science. Science, aac4716. https://doi.org/10.1126/science.aac4716

Orben, A. (2019). A journal club to fix science. *Nature*, *573*, 465–465. https://doi.org/10.1038/d41586-019-02842-8

Ostojić, L., Legg, E. W., Brecht, K. F., Lange, F., Deininger, C., Mendl, M., & Clayton, N. S. (2017). Current desires of conspecific observers affect cache-protection strategies in California scrub-jays and Eurasian jays. *Current Biology*, *27*, R51–R53. https://doi.org/10.1016/j.cub.2016.11.020

Ostojić, L., Shaw, R. C., Cheke, L. G., & Clayton, N. S. (2013). Evidence suggesting that desire-state attribution may govern food sharing in Eurasian jays. *Proceedings of the National Academy of Sciences of the United States of America*, *110*, 4123–4128. https://doi.org/10.1073/pnas.1209926110

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *PLOS Medicine*, *18*, e1003583. https://doi.org/10.1371/journal.pmed.1003583

Paijmans, K. C., Booth, D. J., & Wong, M. Y. L. (2021). Odd one in: Oddity within mixed-species shoals does not affect shoal preference by vagrant tropical damselfish in the presence or absence of a predator. *Ethology*, *127*, 125–134. https://doi.org/10.1111/eth.13110

Papini, M. R. (2003). Comparative psychology. In *Handbook of Research Methods in Experimental Psychology* (pp. 209–240). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470756973.ch10

Pashler, H., & Wagenmakers, E. (2012). Editors' introduction to the special section on replicability in psychological science: a crisis of confidence? *Perspectives on Psychological Science*, *7*, 528–530. https://doi.org/10.1177/1745691612465253

Passos, L., Garcia, G., & Young, R. (2021). Do captive golden mantella frogs recognise wild conspecifics calls? Responses to the playback of captive and wild calls. *Journal of Zoo and Aquarium Research*, *9*, 49–54. https://doi.org/10.19227/jzar.v9i1.476

Patil, P., Peng, R. D., & Leek, J. T. (2016). What should researchers expect when they replicate studies? a statistical view of replicability in psychological science. *Perspectives on Psychological Science*, *11*, 539–544. https://doi.org/10.1177/1745691616646366

Paz-y-Miño C, G., Bond, A. B., Kamil, A. C., & Balda, R. P. (2004). Pinyon jays use transitive inference to predict social dominance. *Nature*, *430*, 778–781. https://doi.org/10.1038/nature02723

Penders, Holbrook, & de Rijcke. (2019). Rinse and repeat: understanding the value of replication across different ways of knowing. *Publications*, *7*, 52. https://doi.org/10.3390/publications7030052

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

Penn, D. C., & Povinelli, D. J. (2013). The comparative delusion. In J. Metcalfe & H. S. Terrace (Eds.), *Agency and Joint Attention* (pp. 62–81). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199988341.003.0004

Pennisi, E. (2020, January 31). *Spider biologist denies suspicions of widespread data fraud in his animal personality research*. Science | AAAS. https://www.sciencemag.org/news/2020/01/spider-biologist-denies-suspicions-widespread-data-fraud-his-animal-personality

Pepperberg, I. M., & Gordon, J. D. (2005). Number comprehension by a grey parrot (*Psittacus erithacus*), including a zero-like concept. *Journal of Comparative Psychology 119*, 197–209. https://doi.org/10.1037/0735-7036.119.2.197

Pereira, F. C., Teixeira, D. L., Boyle, L. A., Pinheiro Machado Filho, L. C., Williams, S. R. O., & Enriquez-Hidalgo, D. (2021). The equipment used in the sf6 technique to estimate methane emissions has no major effect on dairy cow behavior. *Frontiers in Veterinary Science*, *7*, 620810. https://doi.org/10.3389/fvets.2020.620810

Pfungst, O. (1911). *Clever Hans (The Horse of Mr. Von Osten): A Contribution to Experimental Animal and Human Psychology* (1st ed.). Trans. Carl L. Rahn. New York: Holt.

Piefke, T. J., Bonnell, T. R., DeOliveira, G. M., Border, S. E., & Dijkstra, P. D. (2021). Social network stability is impacted by removing a dominant male in replicate dominance hierarchies of a cichlid fish. *Animal Behaviour*, *175*, 7–20. https://doi.org/10.1016/j.anbehav.2021.02.012

Pinto, P., & Hirata, S. (2020). Does size matter? Examining the possible mechanisms of multi-stallion groups in horse societies. *Behavioural Processes*, *181*, 104277. https://doi.org/10.1016/j.beproc.2020.104277

Piper, S. K., Grittner, U., Rex, A., Riedel, N., Fischer, F., Nadon, R., Siegerink, B., & Dirnagl, U. (2019). Exact replication: Foundation of science or game of chance? *PLOS Biology*, *17*, e3000188. https://doi.org/10.1371/journal.pbio.3000188

Polla, E. J., Grueter, C. C., & Smith, C. L. (2018). Asian elephants (*Elephas maximus*) discriminate between familiar and unfamiliar human visual and olfactory cues. *Animal Behavior and Cognition*, *5*(3), 279–291. https://doi.org/10.26451/abc.05.03.03.2018

Popper, K. R. (1962). *Conjectures and refutations: The growth of scientific knowledge* (4. ed. (rev.), repr). Routledge & Kegan Paul.

Poulin-Dubois, D., Rakoczy, H., Burnside, K., Crivello, C., Dörrenberg, S., Edwards, K., Krist, H., Kulke, L., Liszkowski, U., Low, J., Perner, J., Powell, L., Priewasser, B., Rafetseder, E., & Ruffman, T. (2018). Do infants understand false beliefs? We don't know yet – A commentary on Baillargeon, Buttelmann and Southgate's commentary. *Cognitive Development*, *48*, 302–315. https://doi.org/10.1016/j.cogdev.2018.09.005

Povinelli, D. J. (2000). *Folk physics for apes: The chimpanzee's theory of how the world works*. Oxford University Press.

Povinelli, D. J. (2020). Can comparative psychology crack its toughest nut? *Animal Behavior and Cognition*, *7*, 589–652. https://doi.org/10.26451/abc.07.04.09.2020

Povinelli, D. J., & Henley, T. (2020). More rope tricks reveal why more task variants will never lead to strong inferences about higher-order causal reasoning in chimpanzees. *Animal Behavior and Cognition*, *7*, 392–418. https://doi.org/10.26451/abc.07.03.08.2020

Povinelli, D. J., & Vonk, J. (2004). We don't need a microscope to explore the chimpanzee's mind. *Mind & Language*, *19*(1), 1–28. https://doi.org/10.1111/j.1468-0017.2004.00244.x

Prior, H., Schwarz, A., & Güntürkün, O. (2008). Mirror-induced behavior in the magpie (*Pica pica*): evidence of self-recognition. *PLOS Biology*, *6*(8), e202. https://doi.org/10.1371/journal.pbio.0060202

Protzko, J., Krosnick, J., Nelson, L. D., Nosek, B. A., Axt, J., Berent, M., Buttrick, N., DeBell, M., Ebersole, C. R., Lundmark, S., MacInnis, B., O'Donnell, M., Perfecto, H., Pustejovsky, J. E., Roeder, S. S., Walleczek, J., & Schooler, J. (2020). *High Replicability of Newly-Discovered Social-behavioral Findings is Achievable* . PsyArXiv. https://doi.org/10.31234/osf.io/n2a9x

Qu, Z., & Kwok, S. C. (2020). *A meta-analysis on uncertainty monitoring in four non-primate animal species: Pigeons, rats, large-billed crows, and bees* BioRxiv https://doi.org/10.1101/2020.12.03.411082

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

Rasendriya, M. D. (2021, February 25). *How will national regulators in Europe impose fines for missing clinical trial results?* Transparimed. https://www.transparimed.org/single-post/eu-clinical-trial-regulation-ctis-fines

Ravetz, J. R. (1996). *Scientific knowledge and its social problems*. Transaction Publishers.

Redshaw, J., Taylor, A. H., & Suddendorf, T. (2017). Flexible Planning in Ravens? *Trends in Cognitive Sciences*, *21*, 821–822. https://doi.org/10.1016/j.tics.2017.09.001

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

Ribes-Iñesta, E., Hernández, V., & Serrano, M. (2020). Temporal contingencies are dependent on space location: Distal and proximal concurrent water schedules. *Behavioural Processes*, *181*, 104256. https://doi.org/10.1016/j.beproc.2020.104256

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

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

Richter, S. H., Garner, J. P., Zipser, B., Lewejohann, L., Sachser, N., Touma, C., Schindler, B., Chourbaji, S., Brandwein, C., Gass, P., Stipdonk, N. van, Harst, J. van der, Spruijt, B., Võikar, V., Wolfer, D. P., & Würbel, H. (2011). Effect of population heterogenization on the reproducibility of

mouse behavior: a multi-laboratory study. *PLOS ONE*, *6*(, e16461. https://doi.org/10.1371/journal.pone.0016461

Riet, G. ter, Korevaar, D. A., Leenaars, M., Sterk, P. J., Noorden, C. J. F. V., Bouter, L. M., Lutter, R., Elferink, R. P. O., & Hooft, L. (2012). Publication bias in laboratory animal research: a survey on magnitude, drivers, consequences and potential solutions. *PLOS ONE*, *7*, e43404. https://doi.org/10.1371/journal.pone.0043404

Ritchie, S. J., Wiseman, R., & French, C. C. (2012). Failing the future: three unsuccessful attempts to replicate bem's 'retroactive facilitation of recall' effect. *PLOS ONE*, *7*, e33423. https://doi.org/10.1371/journal.pone.0033423

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*, 358–367.

Roche, D. G., Amcoff, M., Morgan, R., Sundin, J., Andreassen, A. H., Finnøen, M. H., Lawrence, M. J., Henderson, E., Norin, T., Speers-Roesch, B., Brown, C., Clark, T. D., Bshary, R., Leung, B., Jutfelt, F., & Binning, S. A. (2020). Behavioural lateralization in a detour test is not repeatable in fishes. *Animal Behaviour*, *167*, 55–64. https://doi.org/10.1016/j.anbehav.2020.06.025

Rohrer, J. M. (2018). Thinking clearly about correlations and causation: graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, *1*, 27–42. https://doi.org/10.1177/2515245917745629

Rohrer, J. M., Tierney, W., Uhlmann, E. L., DeBruine, L. M., Heyman, T., Jones, B. C., Schmukle, S. C., Silberzahn, R., Willén, R. M., Carlsson, R., Lucas, R. E., Vazire, S., Witt, J. K., Zentall, T. R., Chabris, C., & Yarkoni, T. (2018). *Putting the Self in Self-Correction*. PsyArXiv. https://doi.org/10.31234/osf.io/exmb2

Rose, E. M., Mathew, T., Coss, D. A., Lohr, B., & Omland, K. E. (2018). A new statistical method to test equivalence: An application in male and female eastern bluebird song. *Animal Behaviour*, *145*, 77–85. https://doi.org/10.1016/j.anbehav.2018.09.004

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

Rouder, J. N., & Haaf, J. M. (2018). Power, dominance, and constraint: a note on the appeal of different design traditions. *Advances in Methods and Practices in Psychological Science*, *1*(1), 19–26. https://doi.org/10.1177/2515245917745058

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225–237. https://doi.org/10.3758/PBR.16.2.225

Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin*, *57*(5), 416–428. https://doi.org/10.1037/h0042040

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

Scheel, A. M., Schijen, M., & Lakens, D. (2020). *An excess of positive results: Comparing the standard Psychology literature with Registered Reports*. PsyArXiv. https://doi.org/10.31234/osf.io/p6e9c

Scheel, A. M., Tiokhin, L., Isager, P. M., & Lakens, D. (2021). Why hypothesis testers should spend less time testing hypotheses. *Perspectives on Psychological Science*, *16*, 744–755. https://doi.org/10.1177/1745691620966795

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods*, *17*, 551–566. https://doi.org/10.1037/a0029487

Schino, G., Boggiani, L., Mortelliti, A., Pinzaglia, M., & Addessi, E. (2021). Testing the two sides of indirect reciprocity in tufted capuchin monkeys. *Behavioural Processes*, *182*, 104290. https://doi.org/10.1016/j.beproc.2020.104290

Schmitt, V., Pankau, B., & Fischer, J. (2012). Old world monkeys compare to apes in the primate cognition test battery. *PLoS ONE*, *7*, e32024. https://doi.org/10.1371/journal.pone.0032024

Schubiger, M. N. (2019). *Optimising tests of primate cognition: Towards valid species comparisons*. https://doi.org/10.5167/UZH-167963

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

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

Seitz, B. M., Flaim, M. E., & Blaisdell, A. P. (2020). Evidence that novel flavors unconditionally suppress weight gain in the absence of flavor-calorie associations. *Learning & Behavior*, *48*, 351–363. https://doi.org/10.3758/s13420-020-00419-4

Sena, E. S., Worp, H. B. van der, Bath, P. M. W., Howells, D. W., & Macleod, M. R. (2010). Publication bias in reports of animal stroke studies leads to major overstatement of efficacy. *PLOS Biology*, *8*, e1000344. https://doi.org/10.1371/journal.pbio.1000344

Seppelt, R., Beckmann, M., Václavík, T., & Volk, M. (2018). The art of scientific performance. *Trends in Ecology & Evolution*, *33*, 805–809. https://doi.org/10.1016/j.tree.2018.08.003

Sert, N. P. du, Hurst, V., Ahluwalia, A., Alam, S., Avey, M. T., Baker, M., Browne, W. J., Clark, A., Cuthill, I. C., Dirnagl, U., Emerson, M., Garner, P., Holgate, S. T., Howells, D. W., Karp, N. A., Lazic, S. E., Lidster, K., MacCallum, C. J., Macleod, M., … Würbel, H. (2020). The ARRIVE guidelines 2.0: Updated guidelines for reporting animal research. *PLOS Biology*, *18*, e3000410. https://doi.org/10.1371/journal.pbio.3000410

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

Shettleworth, S. J. (1993). Where is the comparison in comparative cognition?: Alternative research programs. *Psychological Science*, *4*, 179–184. JSTOR.

Shettleworth, S. J. (2009). The evolution of comparative cognition: Is the snark still a boojum? *Behavioural Processes*, *80*, 210–217. https://doi.org/10.1016/j.beproc.2008.09.001

Shettleworth, S. J. (2010). Clever animals and killjoy explanations in comparative psychology. *Trends in Cognitive Sciences*, *14*, 477–481. https://doi.org/10.1016/j.tics.2010.07.002

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Dalla Rosa, A., Dam, L., Evans, M. H., Flores Cervantes, I., … Nosek, B. A. (2017). Many analysts, one dataset: Making transparent how variations in analytical choices affect results. *Advances in Methods and Practices in Psychological Science*. http://eprints.nottingham.ac.uk/48166/

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology. *Psychological Science*, *22*, 1359–1366. https://doi.org/10.1177/0956797611417632

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, *9*, 76–80. https://doi.org/10.1177/1745691613514755

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

Simons, K. (2008). The misused impact factor. *Science*. https://www.science.org/doi/abs/10.1126/science.1165316

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*, 534–547. https://doi.org/10.1037/a0033242

Skinner, B. F. (1938). *The behavior of organisms: An experimental analysis* (p. 457). Appleton-Century.

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

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

Smaldino, P. E. (2016). Not even wrong: Imprecision perpetuates the illusion of understanding at the cost of actual understanding. *Behavioral and Brain Sciences*, *39*, e163. https://doi.org/10.1017/S0140525X1500151X

Smaldino, P. E. (2017). Models are stupid, and we need more of them. In R. R. Vallacher, S. J. Read, & A. Nowak (Eds.), *Computational Social Psychology* (1st ed., pp. 311–331). Routledge. https://doi.org/10.4324/9781315173726-14

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, J. D., Couchman, J. J., & Beran, M. J. (2012). The highs and lows of theoretical interpretation in animal-metacognition research. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1594), 1297–1309. https://doi.org/10.1098/rstb.2011.0366

Smith, J. D., Couchman, J. J., & Beran, M. J. (2014). Animal metacognition: a tale of two comparative psychologies. *Journal of Comparative Psychology*, *128*, 115–131. https://doi.org/10.1037/a0033105

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

Smith, N. C. (1970). Replication studies: A neglected aspect of psychological research. *American Psychologist*, *25*, 970–975. https://doi.org/10.1037/h0029774

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

Sober, E. (2005). Comparative psychology meets evolutionary biology: Morgan's canon and cladistic parsimony. *Thinking with Animals: New Perspectives on Anthropomorphism*, 85–99.

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

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

Soto, C. J. (2019). How replicable are links between personality traits and consequential life outcomes? the life outcomes of personality replication project. *Psychological Science*, *30*, 711–727. https://doi.org/10.1177/0956797619831612

Stanley, T. D., Carter, E. C., & Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin*, *144*, 1325–1346. https://doi.org/10.1037/bul0000169

Stapel, D. (2014). *Faking Science: A True Story of Academic Fraud* (Translated by Nicholas J. L. Brown). Self-published. https://errorstatistics.files.wordpress.com/2014/12/fakingscience-20141214.pdf

Starzak, T. B., & Gray, R. D. (2021). Towards ending the animal cognition war: A three-dimensional model of causal cognition. *Biology & Philosophy*, *36*, 9. https://doi.org/10.1007/s10539-021-09779-1

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *11*, 702–712. https://doi.org/10.1177/1745691616658637

Stegenga, J. (2011). Is meta-analysis the platinum standard of evidence? *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *42*(4), 497–507. https://doi.org/10.1016/j.shpsc.2011.07.003

Stengers, I. (2000). *The invention of modern science*. University of Minnesota Press.

Stengers, I., & Muecke, S. (2018). *Another science is possible: A manifesto for slow science* (English edition). Polity.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association*, *54*, 30–34. https://doi.org/10.1080/01621459.1959.10501497

Sterne, J. A. C., Savović, J., Page, M. J., Elbers, R. G., Blencowe, N. S., Boutron, I., Cates, C. J., Cheng, H.-Y., Corbett, M. S., Eldridge, S. M., Emberson, J. R., Hernán, M. A., Hopewell, S., Hróbjartsson, A., Junqueira, D. R., Jüni, P., Kirkham, J. J., Lasserson, T., Li, T., … Higgins, J. P. T. (2019). RoB 2: A revised tool for assessing risk of bias in randomised trials. *BMJ*, *366*, l4898. https://doi.org/10.1136/bmj.l4898

Sterne, J. A. C., Sutton, A. J., Ioannidis, J. P. A., Terrin, N., Jones, D. R., Lau, J., Carpenter, J., Rücker, G., Harbord, R. M., Schmid, C. H., Tetzlaff, J., Deeks, J. J., Peters, J., Macaskill, P., Schwarzer, G., Duval, S., Altman, D. G., Moher, D., & Higgins, J. P. T. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ*, *343*. https://doi.org/10.1136/bmj.d4002

Stevens, A., Doneley, R., Cogny, A., & Phillips, C. J. C. (2021). The effects of environmental enrichment on the behaviour of cockatiels (Nymphicus hollandicus) in aviaries. *Applied Animal Behaviour Science*, *235*, 105154. https://doi.org/10.1016/j.applanim.2020.105154

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

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

Suddendorf, T., & Corballis, M. C. (2008). New evidence for animal foresight? *Animal Behaviour*, *75*, e1–e3. https://doi.org/10.1016/j.anbehav.2008.01.006

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

Takola, E., Krause, E. T., Müller, C., & Schielzeth, H. (2021). Novelty at second glance: A critical appraisal of the novel object paradigm based on meta-analysis. *Animal Behaviour*, *180*, 123–142. https://doi.org/10.1016/j.anbehav.2021.07.018

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

Taylor, A. H. (2014). Corvid cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *5*, 361–372. https://doi.org/10.1002/wcs.1286

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

Tecwyn, E. C. (2021). Doing reliable research in comparative psychology: Challenges and proposals for improvement. *Journal of Comparative Psychology*, *135*(3), 291–301. https://doi.org/10.1037/com0000291

Thom, J. M., & Clayton, N. S. (2013). Re-caching by Western Scrub-Jays (Aphelocoma californica) Cannot Be Attributed to Stress. *PLoS ONE*, *8*, e52936. https://doi.org/10.1371/journal.pone.0052936

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

Tornick, J. K., Rushia, S. N., & Gibson, B. M. (2016). Clark's nutcrackers (*Nucifraga columbiana*) are sensitive to distance, but not lighting when caching in the presence of a conspecific. *Behavioural Processes*, *123*, 125–133. https://doi.org/10.1016/j.beproc.2015.10.023

Troisi, C. A., Cooke, A. C., Davidson, G. L., de la Hera, I., Reichert, M. S., & Quinn, J. L. (2021). No evidence for cross-contextual consistency in spatial cognition or behavioral flexibility in a passerine. *Animal Behavior and Cognition*, *8*, 446–461. https://doi.org/10.26451/abc.08.03.08.2021

Tunç, D. U., Tunç, M. N., & Lakens, D. (2021). *The Epistemic and Pragmatic Function of Dichotomous Claims Based on Statistical Hypothesis Tests*. PsyArXiv. https://doi.org/10.31234/osf.io/af9by

Tuyttens, F. A. M., de Graaf, S., Heerkens, J. L. T., Jacobs, L., Nalon, E., Ott, S., Stadig, L., Van Laer, E., & Ampe, B. (2014). Observer bias in animal behaviour research: Can we believe what we score, if we score what we believe? *Animal Behaviour*, *90*, 273–280. https://doi.org/10.1016/j.anbehav.2014.02.007

Tuyttens, F. A. M., Stadig, L., Heerkens, J. L. T., Van laer, E., Buijs, S., & Ampe, B. (2016). Opinion of applied ethologists on expectation bias, blinding observers and other debiasing techniques. *Applied Animal Behaviour Science*, *181*, 27–33. https://doi.org/10.1016/j.applanim.2016.04.019

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology*, *5*, 207–232. https://doi.org/10.1016/0010-0285(73)90033-9

Tysall, E. E., Pembury Smith, M. Q. R., & Gilman, R. T. (2020). Preregistered report: The effects of marking methodology on mate choice in *Drosophila melanogaster*. *Animal Behavior and Cognition*, *7*, 492–504. https://doi.org/10.26451/abc.07.04.02.2020

van Dalen, H. P., & Henkens, K. (2012). Intended and unintended consequences of a publish-or-perish culture: A worldwide survey. *Journal of the American Society for Information Science and Technology*, *63*, 1282–1293. https://doi.org/10.1002/asi.22636

van der Vaart, E., & Hemelrijk, C. K. (2014). 'Theory of mind' in animals: Ways to make progress. *Synthese*, *191*, 335–354. https://doi.org/10.1007/s11229-012-0170-3

van der Vaart, E., Verbrugge, R., & Hemelrijk, C. K. (2012). Corvid Re-Caching without 'Theory of Mind': A Model. *PLOS ONE*, *7*, e32904. https://doi.org/10.1371/journal.pone.0032904

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

Vasishth, S., Mertzen, D., Jäger, L. A., & Gelman, A. (2018). The statistical significance filter leads to overoptimistic expectations of replicability. *Journal of Memory and Language*, *103*, 151–175. https://doi.org/10.1016/j.jml.2018.07.004

Vater, C., Gray, R., & Holcombe, A. O. (2021). A critical systematic review of the Neurotracker perceptual-cognitive training tool. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-021-01892-2

Vazire, S. (2018). Implications of the credibility revolution for productivity, creativity, and progress. *perspectives on psychological science*, *13*, 411–417. https://doi.org/10.1177/1745691617751884

Vazire, S., & Holcombe, A. O. (2021). Where are the Self-Correcting Mechanisms in Science? *Review of General Psychology*, 10892680211033912. https://doi.org/10.1177/10892680211033912

Vernouillet, A., Clary, D., & Kelly, D. M. (2021). Highly social pinyon jays, but not less social Clark's nutcrackers, modify their food-storing behaviour when observed by a heterospecific. *BioRxiv*, 2021.02.28.433225. https://doi.org/10.1101/2021.02.28.433225

Vernouillet, A., & Kelly, D. M. (2020). Individual exploratory responses are not repeatable across time or context for four species of food-storing corvid. *Scientific Reports*, *10*, 1–11. https://doi.org/10.1038/s41598-019-56138-y

Viglione, G. (2020). 'Avalanche' of spider-paper retractions shakes behavioural-ecology community. *Nature*, *578*(7794), 199–200. https://doi.org/10.1038/d41586-020-00287-y

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

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

Voelkl, B., & Würbel, H. (2016). Reproducibility crisis: are we ignoring reaction norms? *Trends in Pharmacological Sciences*, *37*, 509–510. https://doi.org/10.1016/j.tips.2016.05.003

Voelkl, B., & Würbel, H. (2019). *A Reaction Norm Perspective on Reproducibility* [Preprint]. Developmental Biology. https://doi.org/10.1101/510941

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

Vonk, J. (2016). Advances in Animal Cognition. *Behavioral Sciences*, *6*(4). https://doi.org/10.3390/bs6040027

Vonk, J. (2018). Are chimpanzees “stuck” on their “selves” in video? *Learning & Behavior*, *46*, 227–228. https://doi.org/10.3758/s13420-018-0328-z

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

Vonk, J. (2019b). Emotional contagion or sensitivity to behavior in ravens? *Proceedings of the National Academy of Sciences*, 201909864. https://doi.org/10.1073/pnas.1909864116

Vonk, J. (2021). The journey in comparative psychology matters more than the destination. *Journal of Comparative Psychology*, *135*, 156–167. https://doi.org/10.1037/com0000279

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

Vonk, J., & Shackelford, T. K. (2012). *Comparative Evolutionary Psychology: A United Discipline for the Study of Evolved Traits*. The Oxford Handbook of Comparative Evolutionary Psychology. https://doi.org/10.1093/oxfordhb/9780199738182.013.0029

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology*, *100*, 426–432. https://doi.org/10.1037/a0022790

Wahlsten, D., Metten, P., Phillips, T. J., Boehm, S. L., Burkhart-Kasch, S., Dorow, J., Doerksen, S., Downing, C., Fogarty, J., Rodd-Henricks, K., Hen, R., McKinnon, C. S., Merrill, C. M., Nolte, C., Schalomon, M., Schlumbohm, J. P., Sibert, J. R., Wenger, C. D., Dudek, B. C., & Crabbe, J. C. (2003). Different data from different labs: Lessons from studies of gene-environment interaction. *Journal of Neurobiology*, *54*, 283–311. https://doi.org/10.1002/neu.10173

Wallace, E. K., Altschul, D., Körfer, K., Benti, B., Kaeser, A., Lambeth, S., Waller, B. M., & Slocombe, K. E. (2017). Is music enriching for group-housed captive chimpanzees (*Pan troglodytes*)? *PLOS ONE*, *12*, e0172672. https://doi.org/10.1371/journal.pone.0172672

Waller, B. M., Warmelink, L., Liebal, K., Micheletta, J., & Slocombe, K. E. (2013). Pseudoreplication: A widespread problem in primate communication research. *Animal Behaviour*, *86*, 483–488. https://doi.org/10.1016/j.anbehav.2013.05.038

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, *12*, 129–140. https://doi.org/10.1080/17470216008416717

Wason, P. C. (1968). Reasoning about a Rule. *Quarterly Journal of Experimental Psychology*, *20*(3), 273–281. https://doi.org/10.1080/14640746808400161

Wilson, V. A. D., Kade, C., & Fischer, J. (2021). Testing the relationship between looking time and choice preference in long-tailed macaques. *Animal Behavior and Cognition*, *8*, 351–375. https://doi.org/10.26451/abc.08.03.03.2021

Wu, Y., Petrosky, A. L., Hazzi, N. A., Woodward, R. L., & Sandoval, L. (2021). The role of learning, acoustic similarity and phylogenetic relatedness in the recognition of distress calls in birds. *Animal Behaviour*, *175*, 111–121. https://doi.org/10.1016/j.anbehav.2021.02.015

Würbel, H. (2000). Behaviour and the standardization fallacy. *Nature Genetics*, *26*(3), 263–263. https://doi.org/10.1038/81541

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

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

Wynne, C. D. L. (2004). Fair refusal by capuchin monkeys. *Nature*, *428*, 140–140. https://doi.org/10.1038/428140a

Yang, C., Tsedan, G., Fan, Q., Wang, S., Wang, Z., Chang, S., & Hou, F. (2021). Behavioral patterns of yaks (*Bos grunniens*) grazing on alpine shrub meadows of the Qinghai-Tibetan Plateau. *Applied Animal Behaviour Science*, *234*, 105182. https://doi.org/10.1016/j.applanim.2020.105182

Yarkoni, T. (2018, October 2). No, it's not The Incentives—It's you. *[Citation Needed]*. https://www.talyarkoni.org/blog/2018/10/02/no-its-not-the-incentives-its-you/

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

Yeatman, F. R., & Hirsch, J. (1971). Attempted replication of, and selective breeding for, instrumental conditioning of Drosophila melanogaster. *Animal Behaviour*, *19*, 454–462. https://doi.org/10.1016/s0003-3472(71)80098-2

Yocom, A. M., & Boysen, S. T. (2011). Comprehension of functional support by enculturated chimpanzees Pan troglodytes. *Current Zoology*, *57*, 429–440. https://doi.org/10.1093/czoolo/57.4.429

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

Zwaan, R. A., Pecher, D., Paolacci, G., Bouwmeester, S., Verkoeijen, P., Dijkstra, K., & Zeelenberg, R. (2018). Participant Nonnaiveté and the reproducibility of cognitive psychology. *Psychonomic Bulletin & Review*, *25*, 1968–1972. https://doi.org/10.3758/s13423-017-1348-y

Appendix A – Reference List of Studies Included Chapter 6

Amici, F., Cacchione, T., & Bueno-Guerra, N. (2017). Understanding of object properties by sloth bears, *Melursus ursinus ursinus*. *Animal Behaviour*, *134*, 217–222. https://doi.org/10.1016/j.anbehav.2017.10.028

Anderson, M. R. (2012). Comprehension of object permanence and single transposition in gibbons. *Behaviour*, *149*, 441–459. https://doi.org/10.1163/156853912X639769

Auersperg, A. M. I., Gajdon, G. K., & Huber, L. (2009). Kea (*Nestor notabilis*) consider spatial relationships between objects in the support problem. *Biology Letters*, *5*, 455–458. https://doi.org/10.1098/rsbl.2009.0114

Auersperg, A. M. I., Gajdon, G. K., & Huber, L. (2010). Kea, Nestor notabilis, produce dynamic relationships between objects in a second-order tool use task. *Animal Behaviour*, *80*, 783–789. https://doi.org/10.1016/j.anbehav.2010.08.007

Auersperg, A. M. I., Huber, L., & Gajdon, G. K. (2011). Navigating a tool end in a specific direction: Stick-tool use in kea (*Nestor notabilis*). *Biology Letters*, *7*, 825–828. https://doi.org/10.1098/rsbl.2011.0388

Auersperg, A. M. I., von Bayern, A. M. P., Gajdon, G. K., Huber, L., & Kacelnik, A. (2011). Flexibility in problem solving and tool use of kea and new caledonian crows in a multi access box paradigm. *PLoS ONE*, *6*, Article e20231. https://doi.org/10.1371/journal.pone.0020231

Bailey, I. E., Morgan, K. V., Bertin, M., Meddle, S. L., & Healy, S. D. (2014). Physical cognition: Birds learn the structural efficacy of nest material. *Proceedings of the Royal Society B: Biological Sciences*, *281*, 20133225. https://doi.org/10.1098/rspb.2013.3225

Bird, C. D., & Emery, N. J. (2009). Insightful problem solving and creative tool modification by captive nontool-using rooks. *Proceedings of the National Academy of Sciences*, *106*, 10370–10375. https://doi.org/10.1073/pnas.0901008106

Briefer, E. F., Haque, S., Baciadonna, L., & McElligott, A. G. (2014). Goats excel at learning and remembering a highly novel cognitive task. *Frontiers in Zoology*, *11*, 20. https://doi.org/10.1186/1742-9994-11-20

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*, 207–214. https://doi.org/10.1037/com0000142

Chappell, J., & Kacelnik, A. (2004). Selection of tool diameter by New Caledonian crows *Corvus moneduloides*. *Animal Cognition*, *7*, 121–127. https://doi.org/10.1007/s10071-003-0202-y

Cheke, L. G., Bird, C. D., & Clayton, N. S. (2011). Tool-use and instrumental learning in the Eurasian jay (*Garrulus glandarius*). *Animal Cognition*, *14*, 441–455. https://doi.org/10.1007/s10071-011-0379-4

Cook, R. G., & Fowler, C. (2014). "Insight" in pigeons: Absence of means–end processing in displacement tests. *Animal Cognition*, *17*, 207–220. https://doi.org/10.1007/s10071-013-0653-8

Danel, S., von Bayern, A. M. P., & Osiurak, F. (2019). Ground-hornbills (*Bucorvus*) show means-end understanding in a horizontal two-string discrimination task. *Journal of Ethology*, *37*, 117–122. https://doi.org/10.1007/s10164-018-0565-9

Duranton, C., Rödel, H. G., Bedossa, T., & Belkhir, S. (2015). Inverse sex effects on performance of domestic dogs (*Canis familiaris*) in a repeated problem-solving task. *Journal of Comparative Psychology*, *129*, 84–87. https://doi.org/10.1037/a0037825

Fiset, S., & Plourde, V. (2013). Object permanence in domestic dogs (*Canis lupus familiaris*) and gray wolves (*Canis lupus*). *Journal of Comparative Psychology, 127*, 115–127. https://doi.org/10.1037/a0030595

Gaycken, J., Picken, D. J., Pike, T. W., Burman, O. H. P., & Wilkinson, A. (2019). Mechanisms underlying string-pulling behaviour in green-winged macaws. *Behaviour*, *156*, 619–631. https://doi.org/10.1163/1568539X-00003520

Gajdon, G. K., Ortner, T. M., Wolf, C. C., & Huber, L. (2013). How to solve a mechanical problem: The relevance of visible and unobservable functionality for kea. *Animal Cognition*, *16*, 483–492. https://doi.org/10.1007/s10071-012-0588-5

Girndt, A., Meier, T., & Call, J. (2008). Task constraints mask great apes' ability to solve the trap-table task. *Journal of Experimental Psychology: Animal Behavior Processes*, *34*, 54–62. https://doi.org/10.1037/0097-7403.34.1.54

Helme, A. E., Call, J., Clayton, N. S., & Emery, N. J. (2006). What do bonobos (*Pan paniscus*) understand about physical contact? *Journal of Comparative Psychology*, *120*, 294–302. https://doi.org/10.1037/0735-7036.120.3.294

Helme, A. E., Clayton, N. S., & Emery, N. J. (2006). What do rooks (*Corvus frugilegus*) understand about physical contact? *Journal of Comparative Psychology*, *120*, 288–293. https://doi.org/10.1037/0735-7036.120.3.288

Hoffmann, A., Rüttler, V., & Nieder, A. (2011). Ontogeny of object permanence and object tracking in the carrion crow, *Corvus corone*. *Animal Behaviour*, *82*, 359–367. https://doi.org/10.1016/j.anbehav.2011.05.012

Hofmann, M. M., Cheke, L. G., & Clayton, N. S. (2016). Western scrub-jays (*Aphelocoma californica*) solve multiple-string problems by the spatial relation of string and reward. *Animal Cognition*, *19*, 1103–1114. https://doi.org/10.1007/s10071-016-1018-x

Horner, V., & Whiten, A. (2007). Learning from others' mistakes? Limits on understanding a trap-tube task by young chimpanzees (*Pan troglodytes*) and children (*Homo sapiens*). *Journal of Comparative Psychology*, *121*, 12–21. https://doi.org/10.1037/0735-7036.121.1.12

Joly, M., Micheletta, J., De Marco, A., Langermans, J. A., Sterck, E. H. M., & Waller, B. M. (2017). Comparing physical and social cognitive skills in macaque species with different degrees of social tolerance. *Proceedings of the Royal Society B: Biological Sciences*, *284*, 20162738. https://doi.org/10.1098/rspb.2016.2738

Kittler, K., Kappeler, P. M., & Fichtel, C. (2018). Instrumental problem-solving abilities in three lemur species (*Microcebus murinus, Varecia variegata*, and *Lemur catta*). *Journal of Comparative Psychology*, *132*, 306–314. https://doi.org/10.1037/com0000113

Knaebe, B., Taylor, A. H., Miller, R., & Gray, R. D. (2015). New Caledonian crows (*Corvus moneduloides*) attend to barb presence during pandanus tool manufacture and use. *Behaviour*, *152*, 2107–2125. https://doi.org/10.1163/1568539X-00003316

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

Lacreuse, A., Russell, J. L., Hopkins, W. D., & Herndon, J. G. (2014). Cognitive and motor aging in female chimpanzees. *Neurobiology of Aging*, *35*, 623–632. https://doi.org/10.1016/j.neurobiolaging.2013.08.036

Lampe, M., Bräuer, J., Kaminski, J., & Virányi, Z. (2017). The effects of domestication and ontogeny on cognition in dogs and wolves. *Scientific Reports*, *7*, 11690. https://doi.org/10.1038/s41598-017-12055-6

Liedtke, J., Werdenich, D., Gajdon, G. K., Huber, L., & Wanker, R. (2011). Big brains are not enough: Performance of three parrot species in the trap-tube paradigm. *Animal Cognition*, *14*, 143–149. https://doi.org/10.1007/s10071-010-0347-4

Martin-Ordas, G., Call, J., & Colmenares, F. (2008). Tubes, tables and traps: Great apes solve two functionally equivalent trap tasks but show no evidence of transfer across tasks. *Animal Cognition*, *11*, 423–430. https://doi.org/10.1007/s10071-007-0132-1

Mulcahy, N. J., & Call, J. (2006). How great apes perform on a modified trap-tube task. *Animal Cognition*, *9*, 193–199. https://doi.org/10.1007/s10071-006-0019-6

Mulcahy, N. J., & Schubiger, M. N. (2014). Can orangutans (*Pongo abelii*) infer tool functionality? *Animal Cognition*, *17*, 657–669. https://doi.org/10.1007/s10071-013-0697-9

Mulcahy, N. J., Schubiger, M. N., & Suddendorf, T. (2013). Orangutans (*Pongo pygmaeus and Pongo abelii*) understand connectivity in the skewered grape tool task. *Journal of Comparative Psychology*, *127*, 109–113. https://doi.org/10.1037/a0028621

Müller, C. A., Riemer, S., Range, F., & Huber, L. (2014). Dogs' use of the solidity principle: Revisited. *Animal Cognition*, *17*, 821–825. https://doi.org/10.1007/s10071-013-0709-9

Müller, C. A., Riemer, S., Virányi, Z., Huber, L., & Range, F. (2014). Dogs learn to solve the support problem based on perceptual cues. *Animal Cognition*, *17*, 1071–1080. https://doi.org/10.1007/s10071-014-0739-y

Muth, F., & Healy, S. D. (2014). Zebra finches select nest material appropriate for a building task. *Animal Behaviour*, *90*, 237–244. https://doi.org/10.1016/j.anbehav.2014.02.008

Nawroth, C., Ebersbach, M., & von Borell, E. (2013). A note on pigs' knowledge of hidden objects. *Archives Animal Breeding*, *56*, 861–872. https://doi.org/10.7482/0003-9438-56-086

Nawroth, C., von Borell, E., & Langbein, J. (2015). Object permanence in the dwarf goat (*Capra aegagrus hircus*): Perseveration errors and the tracking of complex movements of hidden objects. *Applied Animal Behaviour Science*, *167*, 20–26. https://doi.org/10.1016/j.applanim.2015.03.010

O'Neill, L., Picaud, A., Maehner, J., Gahr, M., & von Bayern, A. M. P. (2019). Two macaw species can learn to solve an optimised two-trap problem, but without functional causal understanding. *Behaviour*, *156*, 691–720. https://doi.org/10.1163/1568539X-00003521

Painter, M. C., Russell, R. C., & Judge, P. G. (2019). Capuchins (*Sapajus apella*) and squirrel monkeys (*Saimiri sciureus*) fail to attend to the functional spatial relationship between a tool and a reward. *Journal of Comparative Psychology*, *133*, 463–473. https://doi.org/10.1037/com0000179

Pfuhl, G. (2012). Two strings to choose from: Do ravens pull the easier one? *Animal Cognition*, *15*, 549–557. https://doi.org/10.1007/s10071-012-0483-0

Range, F., Hentrup, M., & Virányi, Z. (2011). Dogs are able to solve a means-end task. *Animal Cognition*, *14*, 575–583. https://doi.org/10.1007/s10071-011-0394-5

Range, F., Möslinger, H., & Virányi, Z. (2012). Domestication has not affected the understanding of means-end connections in dogs. *Animal Cognition*, *15*, 597–607. https://doi.org/10.1007/s10071-012-0488-8

Schubiger, M. N., Kissling, A., & Burkart, J. M. (2016). How task format affects cognitive performance: A memory test with two species of New World monkeys. *Animal Behaviour*, *121*, 33–39. https://doi.org/10.1016/j.anbehav.2016.08.005

Seed, A. M., Call, J., Emery, N. J., & Clayton, N. S. (2009). Chimpanzees solve the trap problem when the confound of tool-use is removed. *Journal of Experimental Psychology. Animal Behavior Processes*, *35*, 23–34. https://doi.org/10.1037/a0012925

Seed, A. M., Tebbich, S., Emery, N. J., & Clayton, N. S. (2006). Investigating physical cognition in rooks, *Corvus frugilegus*. *Current Biology*, *16*, 697–701. https://doi.org/10.1016/j.cub.2006.02.066

Seed, A., Seddon, E., Greene, B., & Call, J. (2012). Chimpanzee 'folk physics': Bringing failures into focus. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*, 2743–2752. https://doi.org/10.1098/rstb.2012.0222

St Clair, J. J. H., & Rutz, C. (2013). New Caledonian crows attend to multiple functional properties of complex tools. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *368*, 20120415. https://doi.org/10.1098/rstb.2012.0415

Taylor, A., Roberts, R., Hunt, G., & Gray, R. (2009). Causal reasoning in New Caledonian crows: Ruling out spatial analogies and sampling error. *Communicative & Integrative Biology*, *2*, 311–312. https://doi.org/10.4161/cib.2.4.8224

Taylor, A. H, Hunt, G. R., Medina, F. S., & Gray, R. D. (2009). Do New Caledonian crows solve physical problems through causal reasoning? *Proceedings of the Royal Society B: Biological Sciences*, *276*, 247–254. https://doi.org/10.1098/rspb.2008.1107

Taylor, A. H., Hunt, G. R., Holzhaider, J. C., & Gray, R. D. (2007). Spontaneous metatool use by new Caledonian crows. *Current Biology*, *17*, 1504–1507. https://doi.org/10.1016/j.cub.2007.07.057

Tebbich, S., & Bshary, R. (2004). Cognitive abilities related to tool use in the woodpecker finch, *Cactospiza pallida*. *Animal Behaviour*, *67*, 689–697. https://doi.org/10.1016/j.anbehav.2003.08.003

Tebbich, S., Seed, A. M., Emery, N. J., & Clayton, N. S. (2007). Non-tool-using rooks, *Corvus frugilegus*, solve the trap-tube problem. *Animal Cognition*, *10*, 225–231. https://doi.org/10.1007/s10071-006-0061-4

Tecwyn, E. C., Thorpe, S. K. S., & Chappell, J. (2012). What cognitive strategies do orangutans (*Pongo pygmaeus*) use to solve a trial-unique puzzle-tube task incorporating multiple obstacles? *Animal Cognition*, *15*, 121–133. https://doi.org/10.1007/s10071-011-0438-x

Teschke, I., Cartmill, E. A., Stankewitz, S., & Tebbich, S. (2011). Sometimes tool use is not the key: No evidence for cognitive adaptive specializations in tool-using woodpecker finches. *Animal Behaviour*, *82*, 945–956. https://doi.org/10.1016/j.anbehav.2011.07.032

Teschke, I., & Tebbich, S. (2011). Physical cognition and tool-use: Performance of Darwin's finches in the two-trap tube task. *Animal Cognition*, *14*, 555–563. https://doi.org/10.1007/s10071-011-0390-9

Tia, B., Viaro, R., & Fadiga, L. (2018). Tool-use training temporarily enhances cognitive performance in long-tailed macaques (*Macaca fascicularis*). *Animal Cognition*, *21*, 365–378. https://doi.org/10.1007/s10071-018-1173-3

van Horik, J. O., & Emery, N. J. (2016). Transfer of physical understanding in a non-tool-using parrot. *Animal Cognition*, *19*, 1195–1203. https://doi.org/10.1007/s10071-016-1031-0

Visalberghi, E., & Limongelli, L. (1994). Lack of comprehension of cause-effect relations in tool-using capuchin monkeys (*Cebus apella*). *Journal of Comparative Psychology*, *108*, 15–22.

Weir, A. A. S., & Kacelnik, A. (2006). A New Caledonian crow (*Corvus moneduloides*) creatively re-designs tools by bending or unbending aluminum strips. *Animal Cognition*, *9*, 317–334. https://doi.org/10.1007/s10071-006-0052-5

Yocom, A. M., & Boysen, S. T. (2011). Comprehension of functional support by enculturated chimpanzees *Pan troglodytes*. *Current Zoology*, *57*, 429–440. https://doi.org/10.1093/czoolo/57.4.429

Appendix B – Search Table for Systematic Review of Chapter 7

Table A30: Search terms used for the Scopus search of the corvid social cognition literature

Date of Scopus Search	Keywords	Details	N returned	N selected
07/04/2021	Social cognition	TS=(Social cognition) AND TS=(corvid* OR crow* OR rook* OR jay* OR magpie* OR raven* OR jack daw* OR nutcracker* OR chough*) - Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC Timespan=All years	361	78
07/04/2021	social intelligence OR social brain OR "relationship quality" OR machiavellian	TS = (social intelligence OR social brain OR "relationship quality" OR Machiavellian) AND TS=(corvid* OR "crow" OR "crows" OR rook* OR jay* OR magpie* OR raven* OR jackdaw* OR nutcracker* OR chough*) Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC Timespan=All years	233	48
07/04/2021	collaborat* OR cooperat* OR co-operat*	TS = (collaborat* OR cooperat* OR co-operat*) AND TS=(corvid* OR "crow" OR "crows" OR rook* OR jay* OR magpie* OR raven* OR jackdaw* OR nutcracker* OR chough*) Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC Timespan=All years	755	112
08/04/2021	reciproc* OR altruis* OR spite*	TS = (reciproc* OR altruis* OR spite*) AND TS=(corvid* OR "crow" OR "crows" OR rook* OR jay* OR magpie* OR raven* OR jackdaw* OR nutcracker* OR chough*) Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC Timespan=All years	258	32
08/04/2021	inequit* OR "other regarding" OR "other-regarding" OR prosocial* OR fair*	TS = (inequit* OR "other regarding" OR "other-regarding" OR prosocial* OR fair*) AND TS=(corvid* OR "crow" OR "crows" OR rook* OR jay* OR magpie* OR raven* OR jackdaw* OR nutcracker* OR chough*) Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC Timespan=All years	281	20
08/04/2021	hierarch* OR dominan* OR subordinat*	TS = (hierarch* OR dominan* OR subordinat*) AND TS=(corvid* OR "crow" OR "crows" OR rook* OR jay* OR magpie* OR raven* OR jackdaw* OR nutcracker* OR chough*) Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC Timespan=All years	961	84
09/04/2021	transitive inference OR "self recognition" OR "self-recognition" OR "self awareness" OR "self-awareness" OR mirror response OR mirror recognition	TS = ("transitive inference" OR "self recognition" OR "self-recognition" OR "self awareness" OR "self-awareness" OR mirror response OR mirror recognition) AND TS=(corvid* OR "crow" OR "crows" OR rook* OR jay* OR magpie* OR raven* OR jackdaw* OR nutcracker* OR chough*) Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC Timespan=All years	97	34
09/04/2021	empath* OR "emotional contagion" OR "affect sharing" OR social categorisation OR social categorization OR social discrimination OR conspecific categorisation OR social categorization OR social	TS = (empath* OR "emotional contagion" OR "affect sharing" OR social categorisation OR social categorization OR social discrimination OR conspecific categorisation OR social categorization OR social discrimination OR individual recognition) AND TS=(corvid* OR "crow" OR "crows" OR rook* OR jay* OR magpie* OR raven* OR jackdaw* OR nutcracker* OR chough*) Indexes=SCI-EXPANDED, SSCI, A&HCI,	226	43

	discrimination OR individual recognition	CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC Timespan=All years		
09/04/2021	"social learning" OR "socially learn" OR "learn socially" OR imitat* OR emulat* OR observational learning OR observational cognition OR "stimulus enhancement" OR "local enhancement"	TS = ("social learning" OR "socially learn" OR "learn socially" OR imitat* OR emulat* OR observational learning OR observational cognition OR "stimulus enhancement" OR "local enhancement") AND TS=(corvid* OR "crow" OR "crows" OR rook* OR jay* OR magpie* OR raven* OR jackdaw* OR nutcracker* OR chough*) Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC Timespan=All years	241	61
09/04/2021	theory of mind OR mental state attribution OR mental-state attribution OR mental state-attribution OR knowledge attribution OR attribute knowledge OR ascribe knowledge OR perspective taking OR take perspective OR perspective understanding OR understand perspective	TS = (theory of mind OR mental state attribution OR mental-state attribution OR mental state-attribution OR knowledge attribution OR attribute knowledge OR ascribe knowledge OR perspective taking OR take perspective OR perspective understanding OR understand perspective) AND TS=(corvid* OR "crow" OR "crows" OR rook* OR jay* OR magpie* OR raven* OR jackdaw* OR nutcracker* OR chough*) Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC Timespan=All years	263	16
09/04/2021	gaze following OR follow gaze OR communicat* OR gestur* OR "face inversion" OR "goal-directed" OR "goal directed" OR intention* OR "joint attention" OR "shared attention" OR decept*	TS = (gaze following OR follow gaze OR communicat* OR gestur* OR "face inversion" OR "goal-directed" OR "goal directed" OR intention* OR "joint attention" OR "shared attention" OR decept*) AND TS=(corvid* OR "crow" OR "crows" OR rook* OR jay* OR magpie* OR raven* OR jackdaw* OR nutcracker* OR chough*) Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC Timespan=All years	847	51
12/04/2021[1]	other* AND (belief* OR desire* OR perspective* OR know* OR see	TS = (other* AND (belief* OR desire* OR perspective* OR know* OR see)) AND TS=(corvid* OR "crow" OR "crows" OR rook* OR jay* OR magpie* OR raven* OR jackdaw* OR nutcracker* OR chough*) Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC Timespan=All years	1357	60
14/04/2021	targeted helping OR instrumental helping OR social facilitation OR "self-other" OR "self other" OR mentalizing OR mentalising OR "social referencing" OR reputation	TS = (targeted helping OR instrumental helping OR social facilitation OR "self-other" OR "self other" OR mentalizing OR mentalising OR "social referencing" OR reputation) AND TS=(corvid* OR "crow" OR "crows" OR rook* OR jay* OR magpie* OR raven* OR jackdaw* OR nutcracker* OR chough*) Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC Timespan=All years	81	4
14/04/2021[1]	compet* OR agress* OR social comparison OR "image scoring" OR "third party" OR "third-party" OR "social bond" OR social relationship OR social interaction OR delay of gratification OR self control OR self-control OR temporal discounting OR alliance form* OR conciliation OR reconciliation OR mentalizing OR cross-modal OR tit-for-tat OR prisoner's dilemma OR "MC-PC" OR loose string OR object choice OR	TS = (compet* OR agress* OR social comparison OR "image scoring" OR "third party" OR "third-party" OR "social bond" OR social relationship OR social interaction OR delay of gratification OR self control OR self-control OR temporal discounting OR alliance form* OR conciliation OR reconciliation OR mentalizing OR cross-modal OR tit-for-tat OR prisoner's dilemma OR "MC-PC" OR loose string OR object choice OR object-choice OR social cue use) AND TS=(corvid* OR "crow" OR "crows" OR rook* OR jay* OR magpie* OR raven* OR jackdaw* OR nutcracker* OR chough*) Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC Timespan=All years	1725	127

	object-choice OR social cue use			
15/04/2021	playback OR politics OR social support OR affiliat*	TS = (playback OR politics OR social support OR affiliat*) AND TS=(corvid* OR "crow" OR "crows" OR rook* OR jay* OR magpie* OR raven* OR jackdaw* OR nutcracker* OR chough*) Indexes: SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC. Timespan=All years	736	48
15/04/2021	TS = (conspecific OR other) AND TS=(desire OR knowledge OR belief OR perspective OR intent* OR recogni*OR remember OR memory)	TS = (conspecific OR other) AND TS=(desire OR knowledge OR belief OR perspective OR intent* OR recogni*OR remember OR memory) AND TS=(corvid* OR "crow" OR "crows" OR rook* OR jay* OR magpie* OR raven* OR jackdaw* OR nutcracker* OR chough*) Indexes: SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC. Timespan=All years	832	55
15/04/2021	TS = (fission-fusion OR fission fusion OR social intelligence OR relationship quality)	TS = (fission-fusion OR fission fusion OR social intelligence OR relationship quality) AND TS=(corvid* OR "crow" OR "crows" OR rook* OR jay* OR magpie* OR raven* OR jackdaw* OR nutcracker* OR chough*) Timespan: All years. Indexes: SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC. Timespan=All years	306	36

[1]Due to the large number of search results, a journal-based filter was applied to these searches to remove clearly irrelevant studies. The filter was as follows: Refined by: [excluding] WEB OF SCIENCE CATEGORIES: (MANAGEMENT OR ENGINEERING MECHANICAL OR OPTICS OR LITERATURE OR COMPUTER SCIENCE INFORMATION SYSTEMS OR PHYSICS ATOMIC MOLECULAR CHEMICAL OR CELL BIOLOGY OR PSYCHOLOGY EDUCATIONAL OR ENERGY FUELS OR RELIGION OR GEOGRAPHY PHYSICAL OR TOXICOLOGY OR COMPUTER SCIENCE ARTIFICIAL INTELLIGENCE OR PHILOSOPHY OR ASIAN STUDIES OR ENVIRONMENTAL SCIENCES OR RADIOLOGY NUCLEAR MEDICINE MEDICAL IMAGING OR ENGINEERING MULTIDISCIPLINARY OR BIODIVERSITY CONSERVATION OR PSYCHIATRY OR ENTOMOLOGY OR GEOGRAPHY OR ENGINEERING ELECTRICAL ELECTRONIC OR SOCIAL SCIENCES INTERDISCIPLINARY OR LANGUAGE LINGUISTICS OR HISTORY PHILOSOPHY OF SCIENCE OR MATHEMATICS APPLIED OR MEDICINE GENERAL INTERNAL OR PHYSIOLOGY OR EDUCATION EDUCATIONAL RESEARCH OR PHARMACOLOGY PHARMACY OR TROPICAL MEDICINE OR GEOSCIENCES MULTIDISCIPLINARY OR PHYSICS APPLIED OR BIOTECHNOLOGY APPLIED MICROBIOLOGY OR VIROLOGY OR DERMATOLOGY OR LAW OR IMMUNOLOGY OR ENDOCRINOLOGY METABOLISM OR VETERINARY SCIENCES OR LINGUISTICS OR HOSPITALITY LEISURE SPORT TOURISM OR COMPUTER SCIENCE INTERDISCIPLINARY APPLICATIONS OR OPERATIONS RESEARCH MANAGEMENT SCIENCE OR MATHEMATICAL COMPUTATIONAL BIOLOGY OR GENETICS HEREDITY OR TELECOMMUNICATIONS OR MEDICINE RESEARCH EXPERIMENTAL OR MARINE FRESHWATER BIOLOGY OR COMPUTER SCIENCE THEORY METHODS OR NUTRITION DIETETICS OR ENVIRONMENTAL STUDIES OR OCEANOGRAPHY OR BIOCHEMISTRY MOLECULAR BIOLOGY OR FORESTRY OR PATHOLOGY OR PUBLIC ENVIRONMENTAL OCCUPATIONAL HEALTH OR HEALTH CARE SCIENCES SERVICES OR POLITICAL SCIENCE OR GASTROENTEROLOGY HEPATOLOGY OR HUMANITIES MULTIDISCIPLINARY OR ANTHROPOLOGY OR MATERIALS SCIENCE MULTIDISCIPLINARY OR URBAN STUDIES OR PLANT SCIENCES OR PALEONTOLOGY OR AGRONOMY OR HISTORY OR WATER RESOURCES OR COMPUTER SCIENCE SOFTWARE ENGINEERING OR PARASITOLOGY OR AGRICULTURE DAIRY ANIMAL SCIENCE OR CONSTRUCTION BUILDING TECHNOLOGY OR CLINICAL NEUROLOGY OR BUSINESS OR FISHERIES OR PHYSICS MULTIDISCIPLINARY OR COMMUNICATION OR GEOCHEMISTRY GEOPHYSICS OR SURGERY OR ENGINEERING CIVIL OR MECHANICS OR ARCHAEOLOGY OR ENGINEERING INDUSTRIAL OR ONCOLOGY OR INFECTIOUS DISEASES)