ABC 2020, 7(3):419-444
Animal Behavior and Cognition
DOI: https://doi.org/10.26451/abc.07.03.09.2020

Trialling Meta-Research in Comparative Cognition: Claims and Statistical Inference in Animal Physical Cognition

Benjamin G. Farrar[1,6*], Drew M. Altschul[2,3], Julia Fischer[4], Jolene van der Mescht[2,3], Sarah Placì[4], Camille A. Troisi[5], Alizée Vernouillet[1], Nicola S. Clayton[1], and Ljerka Ostojić[6]

[1]Department of Psychology, University of Cambridge, Cambridge, UK
[2]Department of Psychology, The University of Edinburgh, Edinburgh, UK
[3]Scottish Primate Research Group, UK
[4]Cognitive Ethology Laboratory, German Primate Center, Göttingen, Germany
[5]School of Biological, Earth and Environmental Sciences, University College Cork, Cork, Ireland
[6]Institute for Globally Distributed Open Research and Education (IGDORE)

*Corresponding author (Email: bgf22@cam.ac.uk)

Citation – Farrar, B. G., Altschul, D. M., Fischer, J., van der Mescht, J., Placì, S., Troisi, C. A., Vernouillet, 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), 419-444. doi: https://doi.org/10.26451/abc.07.03.09.2020

Abstract – Scientific disciplines face concerns about replicability and statistical inference, and these concerns are also relevant in animal cognition research. This paper presents a first attempt to assess how researchers make and publish claims about animal physical cognition, and the statistical inferences they use to support them. We surveyed 116 published experiments from 63 papers on physical cognition, covering 43 different species. The most common tasks in our sample were trap-tube tasks (14 papers), other tool use tasks (13 papers), means-end understanding and string-pulling tasks (11 papers), object choice and object permanence tasks (9 papers) and access tasks (5 papers). This sample is not representative of the full scope of physical cognition research; however, it does provide data on the types of statistical design and publication decisions researchers have adopted. Across the 116 experiments, the median sample size was 7. Depending on the definitions we used, we estimated that between 44% and 59% of our sample of papers made positive claims about animals' physical cognitive abilities, between 24% and 46% made inconclusive claims, and between 10% and 17% made negative claims. Several failures of animals to pass physical cognition tasks were reported. Although our measures had low inter-observer reliability, these findings show that negative results can and have been published in the field. However, publication bias is still present, and consistent with this, we observed a drop in the frequency of *p*-values above .05. This suggests that some non-significant results have not been published. More promisingly, we found that researchers are likely making many correct statistical inferences at the individual-level. The strength of evidence of statistical effects at the group-level was weaker, and its *p*-value distribution was consistent with some effect sizes being overestimated. Studies such as ours can form part of a wider investigation into statistical reliability in comparative cognition. However, future work should focus on developing the validity and reliability of the measurements they use, and we offer some starting points.

Keywords – Physical cognition, Folk physics, Evidence, Statistical inference, Publication bias

Povinelli's (2000) "Folk Physics for Apes" had a lasting impact on the study of animals' physical cognition. However, in "Folk Physics for Apes", Povinelli also highlighted many of the issues at the forefront of science's replication crisis (see Flis, 2019), such as replication (Open Science Collaboration, 2015; Zwaan et al., 2018), falsification (LeBel et al., 2017), strong inference (Stevens, 2017) and being willing and able to interpret negative results (Lakens, 2017a). These issues focus on researchers' statistical inferences, for example about how they may overestimate the support a significant *p*-value

gives for a given hypothesis (McShane et al., 2019). In contrast to statistical concerns, most criticism in animal physical cognition research has focused on construct validity, i.e., the extent to which tasks that researchers present to animals can provide diagnostic information on the cognitive processes that animals use (e.g., Ghirlanda & Lind, 2017; Hennefield et al., 2018). To date, there has been little assessment of the reliability and robustness of statistical inferences in the field. Motivated by the apparent success of meta-research projects in related fields, we conducted a preliminary assessment of statistical designs and reliability in a sample of animal physical cognition experiments.

Statistical Design

Sample Size and Statistical Biological Unit of Interest

Our 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). Second, small samples may poorly represent the researcher's overall target population, leading to over-generalized claims (Henrich et al., 2010; Hurlbert, 1984). 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 & 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 original work, Megan the chimpanzee was correct on 80/100 trials, (p = .000000000135); no other chimpanzee performed significantly above chance. 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 our 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.

Publication Bias and Statistical Reliability

Our 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 contains 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.

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 are likely some 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 we aimed to examine in animal physical cognition research.

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, 2017b; 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. We 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.

Proportion of Animals "Passing" a Test

Finally, we 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.

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/.

Paper Inclusion

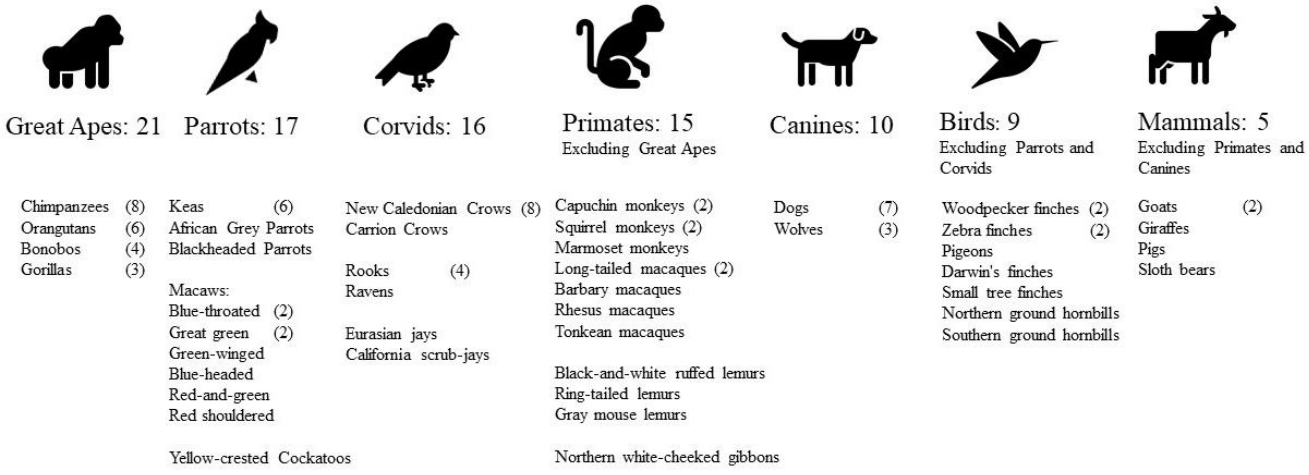
This project attempted to code information from 200 published experiments in animal physical cognition studies. Papers were found using a keyword search in Scopus. Papers with titles, abstracts and/or keywords containing the following keywords: "folk physics" OR "physical cognition" were searched, returning 167 results on 26th November 2019. An error in an earlier search meant we had expected over 600 results from this search, and after realizing that the search returned only 167 results,

we 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 our 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 minimizing the extent our 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 we 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 our 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 Appendix A. There were 45 different species represented across the 63 physical cognition papers that we coded (Figure 1), with chimpanzees (8), New Caledonian crows (8), dogs (7), orangutans (6) and keas (6) being the most common.

Figure 1

The Number of Papers Investigating Each Species or Family in Our Sample of Physical Cognition Papers



Note. The sample consisted of behavioral experiments on captive animal physical cognition and was sampled using a Scopus search for "folk physics" OR "physical cognition," and a variety of trap-tube and trap-table combinations. If a species was in more than one paper, this number is given in brackets after the species; e.g., there were 8 papers that included chimpanzees in the sample, and two papers including goats.

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 chose 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.

Our sample is weakly representative of physical cognition research, and this sampling limits the ability of our analysis to characterize animal physical cognition research in general. Because of our search criteria, trap-tube and tool use studies are over-represented in our sample, some task formats are only represented once (e.g., water raising tasks), and some are not represented (e.g., violation of expectation tasks). Our sample is clustered around topics typically associated with "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. The sample does not include experiments on spatial cognition, or numerical cognition.

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". We 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 were asked to copy 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, one author (BGF) 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 we did not have sufficient information to code the exact *p*-value, e.g., $p < .00001$ ($N = 2$) and $p < .001$ ($N = 2$), we included these as equalities in our analysis. We additionally planned to assess how researchers interpreted non-significant *p*-values; however, we decided that the number of non-significant *p*-values that we coded (16) was insufficient for any robust analysis.

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 our final sample were double coded: BGF coded all of the experiments, and CAT, 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). 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 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 we archived in our working methods, which was that the coders agreed if 50% of the text overlapped. We changed this criterion as the coders often varied in how much text they included, despite focusing on the same claim or inference.

Analyses

We first present descriptive data and visualizations of the following information: the distribution of species and groups from our 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. We then use these data to explore the types of claims and strength of statistical inferences present in the animal physical cognition literature. We 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, or studying zero true effects, i.e., 5% power (for simulation details, see Appendix B). We make 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). We explain this uncertainty in the Discussion section, as seeing the data first can help illustrate the limitations of the analysis.

Results

Coding Reliability

The first two coders agreed on 56 out of 63 (89%) claims from each paper's abstract, Cohen's kappa = .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 $kappa_{weighted}$ = .47. Coders agreed on 45 of 75 (60%) group-level inferences, Cohen's kappa = .60, and 63 of 93 (68%) individual-level inferences, Cohen's kappa =.68. Sample sizes were coded equally in 103 of 116 experiments, Cohen's kappa = .89, and the number of animals passing each test was agreed in 56 of 94, Cohen's kappa = .59. All disagreements were then resolved by a third coder to produce the dataset we used in 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 our *p*-distribution analyses twice. Our primary analysis was performed on the inferences that both coders agreed with, or that the third coder decided on in cases of disagreement. Our 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) our 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 characterize the papers, and iv) typographical errors.

To investigate this further, we 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, we 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 2 visualizes 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 2 shows how these claims were separated (negative leftmost, inconclusive center, positive rightmost). Our 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 2 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. Overall, we feel this level of inter-rater agreement is acceptable for our purpose of categorizing claims. However, perhaps more important was that our original definitions were vague. To assess the effects of vague definitions on the findings, we 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 behavior displayed by the animals. Specifically:

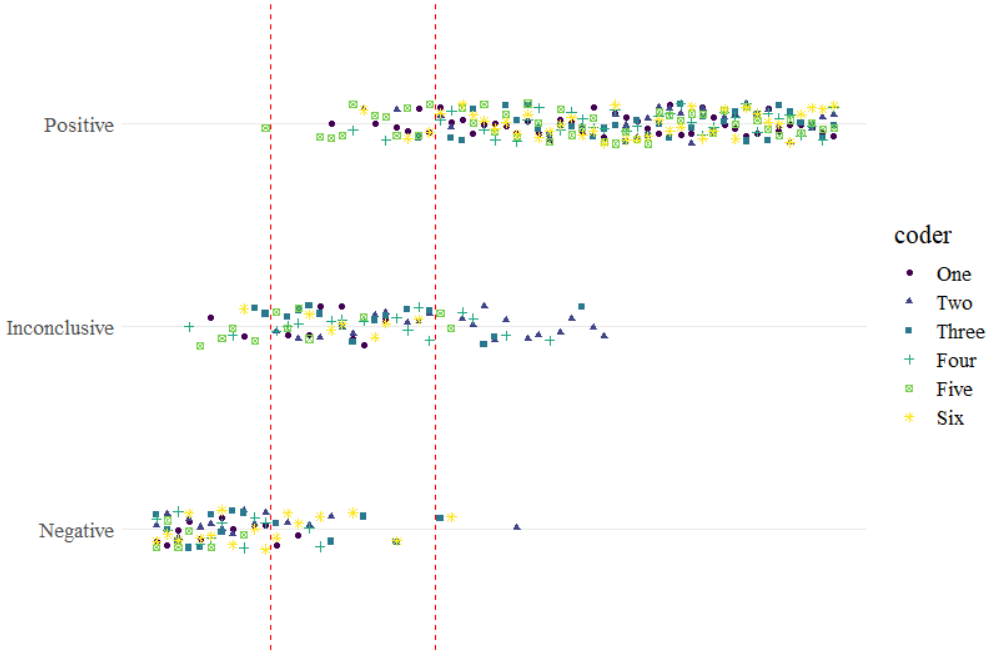
Positive claims will include, but are not limited to, general statements about animals: having a certain cognitive ability; being able to pass a test; any claimed discovery of the processes animals use when passing a test (excluding "lower-order" processes); or confirmation of a general hypothesis. For 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
The animals used a two-step procedure to pass the task, first assessing the causal structure and then selecting the most appropriate tool
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)

Figure 2

Inter-rater Agreement When All 63 Claims Were Re-coded by 6 Coders Using the Original Coding Criteria



Note. 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' (center) when the number of coders coding the paper as either positive/negative or inconclusive was equally split, (i.e., 3 coders in each category).

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

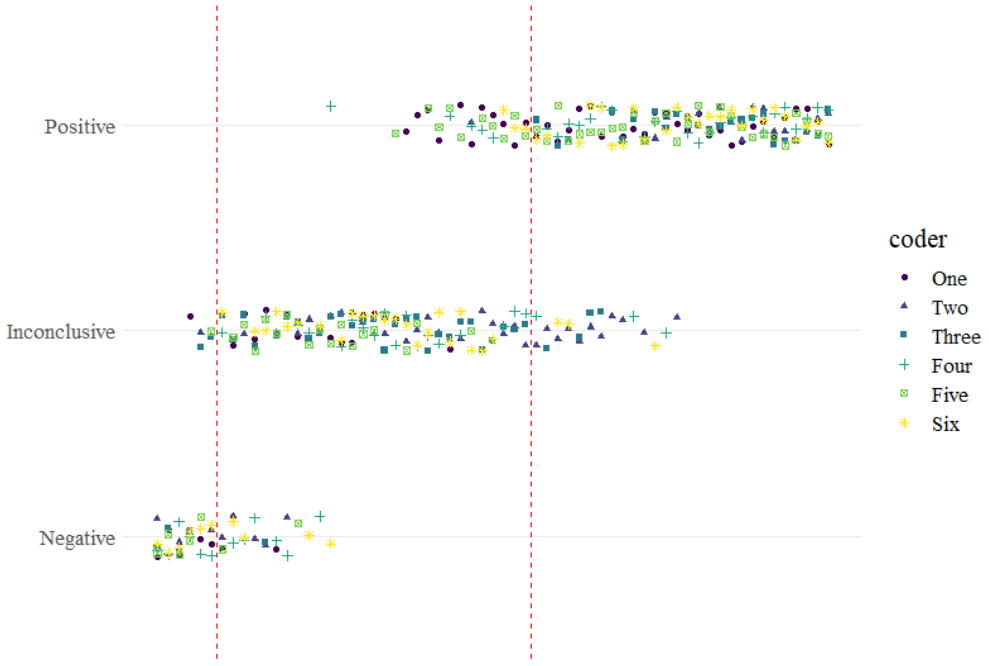
Inconclusive claims will either not make strong epistemic statements but point to task specific confounds or task specific alternative explanations, OR they will report/refer to mixed evidence, i.e., some positive and some negative results OR a failure to confirm or disconfirm a hypothesis. For example:

> 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 3). 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 3

Inter-rater Agreement When All 63 Claims Were Re-coded by 6 Coders Using More Specific Coding Definitions



Note. 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' (center) when the number of coders coding the paper as either positive/negative or inconclusive was equally split, (i.e., 3 coders in each category).

Even though we were unable to increase our 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 we 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 1. Next, we present the results of our analyses.

Table 1

Examples of how Claims were Coded from Our Papers Across Two Rounds of Coding by 6 Individuals

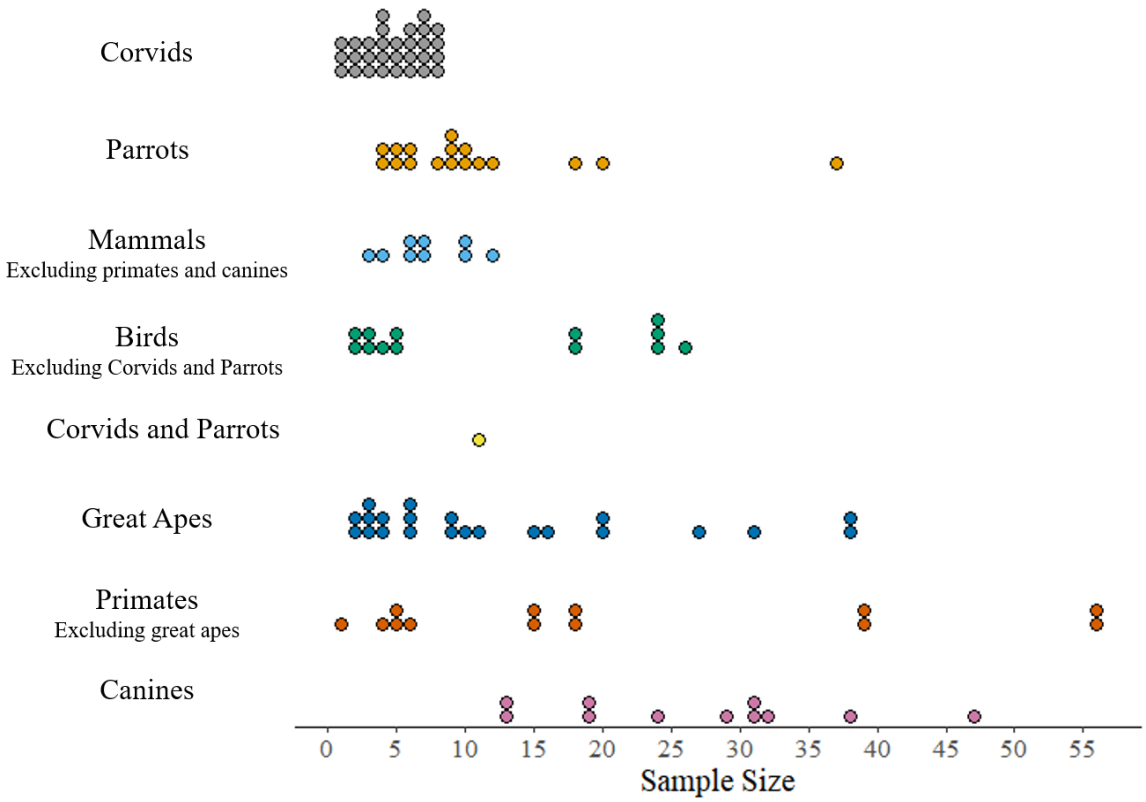
Claim level category	Example
Positive	"The results indicate that kea are capable of assessing the spatial means–end relationships of this problem spontaneously and in a way that is comparable with primates." (Auersperg et al., 2009)
	"Our results showed that tool-use training enhances mean performance 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)

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 4). 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 4

The Distribution of the Total Sample Sizes from Each Experiment of Our Sample of Physical Cognition Experiments



Claims

From our two coding rounds in which six individuals coded each claim, we 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. Table 2 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.

P-values

From the 68 group-level inferences, we had sufficient information to extract 58 exact *p*-values, of which 46 were < .05. From the 88 individual-level inferences, we 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 5, 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 we 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 6.

Table 2

The Number of Each Type of Claim (Positive, Negative or Inconclusive) Made in the Sample of Physical Cognition Papers, by Study Group

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 Excluding 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 and parrots	3; 3	3; 2	1; 2

Note. 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.

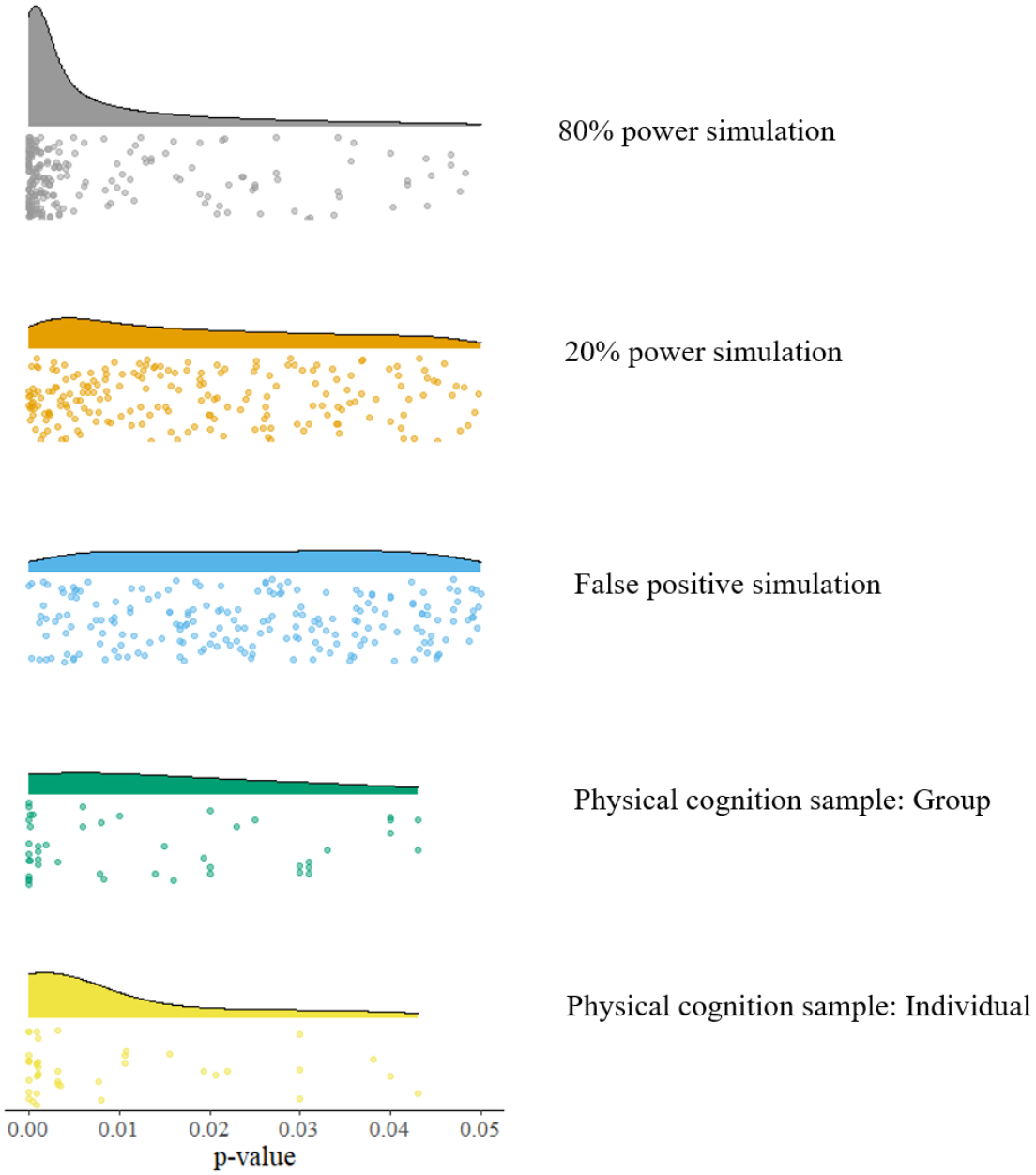
We performed two further exploratory analyses with the group-level *p*-values. The upper panel of Figure 7 displays the p-values as a function of sample size, with a red dashed line at $p = .05$, and the area $.01 < p < .05$ shaded. This graph provides weak evidence for larger sample sizes to produce smaller *p*-values, however some very small *p*-values were still reported from studies with small sample sizes. The lower panel of Figure 7 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.

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, no (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 8 (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 (Figure 8, bottom panel).

Figure 5

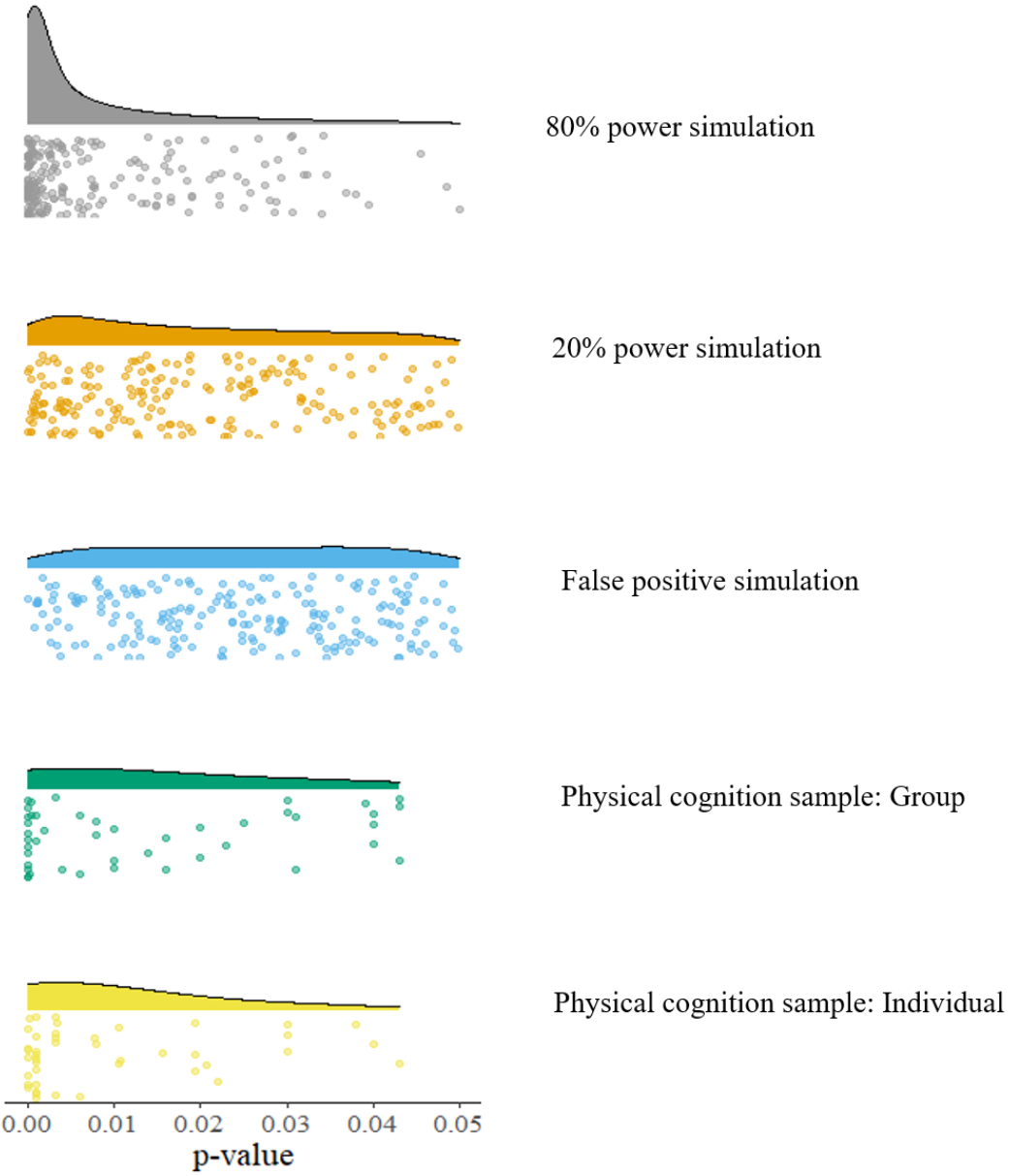
Raincloud Plot of Significant P-Values Observed in Our Physical Cognition Sample and From Simulated Research At Different Statistical Powers



Note. For the simulated research, H1 was always correct for the research performed with 80% power (uppermost plot), 20% power (second panel from top), and H0 was always correct for the false positive simulation (third from top). The fourth panel from top displays the significant group-level Inferences from our sample of physical cognition research (N=46, Fourth Panel) and the bottom panel displays the significant results from 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 6

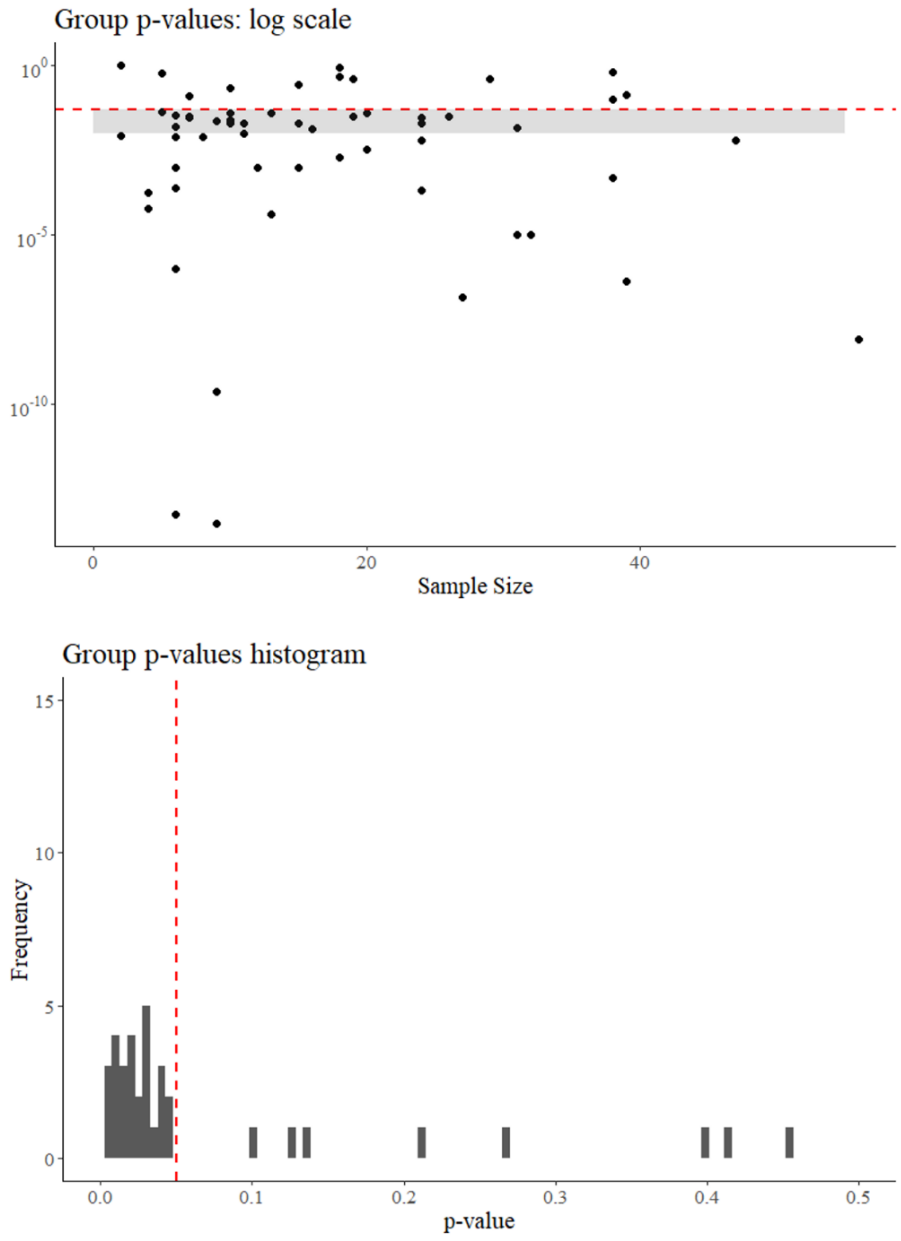
Raincloud Plot of Significant P-Values from the Alternative Statistical Inferences from Our Sample of Physical Cognition Research.



Note. The top three panels display simulated research, and the bottom two the alternative *p*-values from our study. For the simulated research, H1 was always correct for the research performed with 80% power (uppermost plot), 20% power (second panel from top), and H0 was always correct for the false positive simulation (third from top). The fourth panel from top displays the significant group-level Inferences from our sample of physical cognition research (N=46, Fourth Panel) and the bottom panel displays the significant results from 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 7

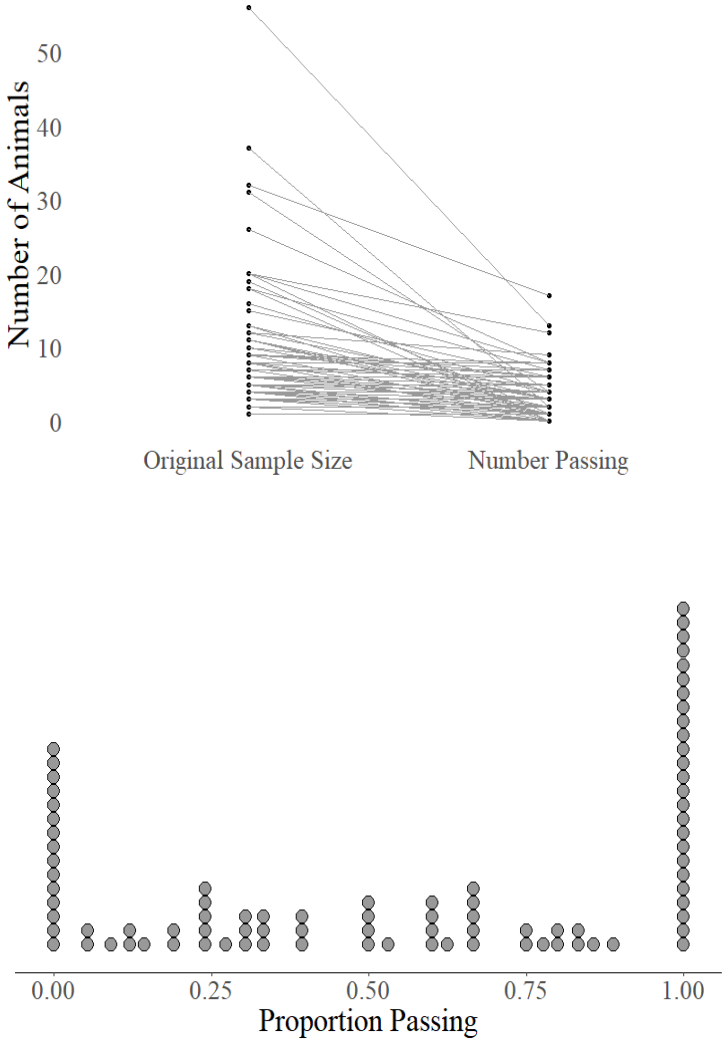
The Distribution of Group p-Values from Our Sample as a Function of Sample Size (Upper Panel), and Frequency (Lower Panel)



Note. 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 we excluded this range for clarity). The x-axis gives the p-values and the y-axis shows the frequency with which they are reported. The red dashed line denotes a *p*-value of .05.

Figure 8

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).



Discussion

In this study, we 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 we 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 our sample). Because of these sampling biases, the exact numbers for many of our 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.

Statistical Design

Sample Size and Biological Unit of Interest

In line with previous reports of sample sizes in animal cognition research and comparative psychology (Craig & Abramson, 2018; Many Primates et al., 2019), we found that many experiments tested fewer than 10 animals, with a median of 7. However, we found sample sizes of greater than 20 for several species, for example, Kittler et al. (2018) tested 57 lemurs of three different species, Duranton et al. (2015) tested 47 dogs, and Joly et al. (2017) tested a total of 39 macaques of four different species. Our 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 key-words 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. We found that, in 13 experiments, between 1% and 25% of the sample passed the test at hand (Figure 8), showing that small sample sizes can still be a concern, even if researchers focus their statistical analyses on the individual.

Publication Bias and Statistical Reliability

Depending on our definition of 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. 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, we 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 our own, 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. Estimating publication bias will therefore require a mix of research methods, such as survey measures, getting direct access to unpublished data or following-up on registered studies (Franco et al., 2014), in addition to the methods we have used here. Such methods will be most effective when they focus on narrow research questions and sufficient data are collected to examine individual research groups, editors, and journals.

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). We could not perform formal *p*-curve analyses with our data because *p*-curve analyses, to be effective, require i) independent data points; whereas we 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.

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

Our 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.

Sampling and Level of Analysis

The generalizability of our 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 we 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 or take a random sample of articles from across animal behavior and animal cognition journals. This will provide more comprehensive data on the statistical design of experiments. However, it is uncertain whether collecting more, and more representative, data is the best strategy for assessing statistical reliability using the methods we trialed here. *P*-distribution analyses, for example, are most valid when used to assess a family of research testing the same or very similar hypotheses (Simonsohn et al., 2014), and publication bias may be best viewed at the level of the individual research group, or journal. To address the prevalence of publication bias and the statistical reliability of animal cognition research, we suggest that researchers focus on generating high quality data across narrow domains, rather than data spread thinly across many different topics.

Definitions and Reliability

The inter-rater reliabilities of our measures were low, specifically when researchers assessed the level of a claim, which we 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, we believe if such measures were used to generate near 100% inter-rater reliabilities, this would reduce the measures' validities (Penders et al., 2019). In particular, we view the claim levels (positive, inconclusive or negative) as somewhat more like a survey measure than rote data extraction. However, it is clear that our definitions can be improved, and this may be achieved through generating more continuous measurement scales through a larger number of ordinal categories, or a 0-100 scale, or through using additional categories, such as splitting inconclusive into categories such "mixed-evidence" and "methodological problems." While the validity and reliability of measures like the ones we used are being explored, researchers may be able to increase the robustness of their conclusions through using heterogeneous methods, for example by using multiple definitions, many coders, and multiverse-type

analyses. Overall, we believe that projects like our own will provide a valuable thread of information in a wider investigation of publication bias and statistical reliability in comparative cognition. However, further research is needed to establish exactly what these projects should measure, and the reliability and validity of the measurements they provide.

Conclusion

The present study analyzed 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 we 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, we observed 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. Finally, we encountered problems with our sample and the reliability of our measurements and discussed how they could be remedied as part of a larger body of research investigating publication bias and statistical reliability in comparative cognition.

Acknowledgments

We would like to thank Mahmoud Elsherif for comments on the manuscript, and Edward Legg for performing a reproducibility check of our code. BGF was supported by the University of Cambridge BBSRC Doctoral Training Programme (BB/M011194/1). DMA was supported by an MRC Mental Health Data Pathfinder award (MC_PC_17209).

Author Contributions

BGF and LO designed the study, with input from all other authors. BGF, DA, JF, JvdM, SP, CAT, AV, and LO collected the data. BGF wrote the original draft, and all authors contributed to interpreting the results and revising the paper.

Conflict of Interest Statement

The authors report no conflicts of interest in the publication of this manuscript.

References

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

Barlow, D. H., & Nock, M. K. (2009). Why can't we be more idiographic in our research? *Perspectives on Psychological Science*, *4*, 19–21. https://doi.org/10.1111/j.1745-6924.2009.01088.x

Button, K. S., A 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

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*, *31*, Retrieved from https://escholarship.org/uc/item/08w0c08s

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

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

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *PLOS ONE*, *5*, Article e10068. https://doi.org/10.1371/journal.pone.0010068

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*, 1–22. https://doi.org/10.26451/abc.07.01.02.2020

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

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

Flis, I. (2019). Psychologists psychologizing scientific psychology: An epistemological reading of the replication crisis. *Theory & Psychology*, *29*, 158–181. https://doi.org/10.1177/0959354319835322

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, *345*, 1502–1505. https://doi.org/10.1126/science.1255484

Ghirlanda, S., & Lind, J. (2017). 'Aesop's fable' experiments demonstrate trial-and-error learning in birds, but no causal understanding. *Animal Behaviour*, *123*, 239–247. https://doi.org/10.1016/j.anbehav.2016.10.029

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*, 61. https://doi.org/10.2307/1164832

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

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*, 29–29. https://doi.org/10.1038/466029a

Herrmann, E., Call, J., Hernàndez-Lloreda, M. V., Hare, B., & Tomasello, M. (2007). Humans have evolved specialized skills of social cognition: The cultural intelligence hypothesis. *Science*, *317*, 1360–1366. https://doi.org/10.1126/science.1146282

Hurlbert, S. H. (1984). Pseudoreplication and the design of ecological field experiments. *Ecological Monographs*, *54*, 187–211. https://doi.org/10.2307/1942661

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

Lakens, D. (2017a). 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. (2017b). *Professors are not elderly: Evaluating the evidential value of two social priming effects through p-curve analyses* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/3m5y9

LeBel, E. P., Berger, D., Campbell, L., & Loving, T. J. (2017). Falsifiability is not optional. *Journal of Personality and Social Psychology*, *113*, 254–261. https://doi.org/10.1037/pspi0000106

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*, 205-220

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

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

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

Nissen, S. B., Magidson, T., Gross, K., & Bergstrom, C. T. (2016). Publication bias and the canonization of false facts. *ELife*, *5*, Article e21451. https://doi.org/10.7554/eLife.21451

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

Penders, B., Holbrook, J. B., & de Rijcke, S. (2019). Rinse and repeat: Understanding the value of replication across different ways of knowing. *Publications*, *7*, 52. https://doi.org/10.3390/publications7030052

Povinelli, D. J. (2000). *Folk physics for apes: The chimpanzee's theory of how the world works*. Oxford University Press.

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. (1956). A case history in scientific method. *American Psychologist*, *11*, 221–233. https://doi.org/10.1037/h0047662

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

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

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

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

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

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. https://doi.org/ 10.1037/0735-7036.108.1.15

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

Appendix A: Journals and references of the papers from which data were extracted

Table A1

The Journals in which Our Sample of Physical Cognition Papers Were Published.

Journal	Number of Papers
Animal Behaviour	7
Animal Cognition	20
Applied Animal Behaviour Science	1
Archives Animal Breeding	1
Behaviour	5
Biology Letters	2
Communicative and Integrative Biology	1
Current Biology	2
Current Zoology	1
Frontiers in Zoology	1
Journal of Comparative Psychology	10
Journal of Ethology	1
Journal of Experimental Psychology: Animal Behavior	2
Neurobiology of Aging	1
Philosophical Transactions of the Royal Society B: Biological Sciences	2
PloS ONE	1
PNAS	1
Proceedings of the Royal Society B: Biological Sciences	3
Scientific Reports	1

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. https://doi.org/10.1037/0735-7036.108.1.15

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: Simulation Details

For Figures 5 and 6, data were simulated from two normal distributions for each of the four sets of simulations:

Power	**Population 1**	**Population 2**
80%	$X \sim N(50, 5)$	$X \sim N(52.02, 5)$
20%	$X \sim N(50, 5)$	$X \sim N(50.81, 5)$
5%	$X \sim N(50, 5)$	$X \sim N(50, 5)$

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 5 and 6 alongside the *p*-values we sampled from the physical cognition literature.

For Figures 5 and 6, the *p*-values we sampled from the physical cognition literature were plotted alongside a mixture model of research at 80% power (1/12 of all research), 20% power (2/3 of all research) and 5% power (1/4 of all research) from the same populations as Figures 5 and 6.