Reporting and interpreting non-significant results in animal cognition research

Farrar, Benjamin G; Vernouillet, Alizée; Garcia-Pelegrin, Elias; Legg, Edward W; Brecht, Katharina F; Lambert, Poppy L; Elsherif, Mahmoud; Francis, Shannon; O'Neill, Laurie; Clayton, Nicola S; ...

Source / Izvornik: **PeerJ – the Journal of Life & Environmental Sciences, 2023, 11**

Journal article, Published version
Rad u časopisu, Objavljena verzija rada (izdavačev PDF)

https://doi.org/10.7717/peer

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:186:770670

Rights / Prava: Attribution 4.0 International/Imenovanje 4.0 međunarodna

Download date / Datum preuzimanja: **2025-01-23**

Repository / Repozitorij:

Repository of the University of Rijeka, Faculty of Humanities and Social Sciences - FHSSRI Repository

Reporting and interpreting non-significant results in animal cognition research

Benjamin G. Farrar[1,2], Alizée Vernouillet[3], Elias Garcia-Pelegrin[1,4], Edward W. Legg[5,6,7], Katharina F. Brecht[8], Poppy J. Lambert[9], Mahmoud Elsherif[10,11], Shannon Francis[12], Laurie O'Neill[12], Nicola S. Clayton[1] and Ljerka Ostojić[5,6,7]

[1] Department of Psychology, University of Cambridge, Cambridge, United Kingdom
[2] Institute for Globally Distributed Open Research and Education (IGDORE), Cambridge, United Kingdom
[3] Department of Experimental Psychology, Universiteit Gent, Gent, Belgium
[4] Department of Psychology, National University of Singapore, Singapore, Singapore
[5] Department of Psychology, Faculty of Humanities and Social Sciences, University of Rijeka, Rijeka, Croatia
[6] Division of Cognitive Sciences, University of Rijeka, Rijeka, Croatia
[7] Centre for Mind and Behaviour, University of Rijeka, Rijeka, Croatia
[8] Institute for Neurobiology, University of Tuebingen, Tuebingen, Germany
[9] Messerli Research Insititute, University of Vienna, Vienna, Austria
[10] Department of Psychology, University of Birmingham, Birmingham, United Kingdom
[11] University of Leicester, Leicester, United Kingdom
[12] Comparative Cognition Research Group, Max Planck Institute for Ornithology, Seewiesen, Germany

ABSTRACT

How statistically non-significant results are reported and interpreted following null hypothesis significance testing is often criticized. This issue is important for animal cognition research because studies in the field are often underpowered to detect theoretically meaningful effect sizes, *i.e.*, often produce non-significant *p*-values even when the null hypothesis is incorrect. Thus, we manually extracted and classified how researchers report and interpret non-significant *p*-values and examined the *p*-value distribution of these non-significant results across published articles in animal cognition and related fields. We found a large amount of heterogeneity in how researchers report statistically non-significant *p*-values in the result sections of articles, and how they interpret them in the titles and abstracts. Reporting of the non-significant results as "No Effect" was common in the titles (84%), abstracts (64%), and results sections (41%) of papers, whereas reporting of the results as "Non-Significant" was less common in the titles (0%) and abstracts (26%), but was present in the results (52%). Discussions of effect sizes were rare (<5% of articles). A *p*-value distribution analysis was consistent with research being performed with low power of statistical tests to detect effect sizes of interest. These findings suggest that researchers in animal cognition should pay close attention to the evidence used to support claims of absence of effects in the literature, and—in their own work—report statistically non-significant results clearly and formally correct, as well as use more formal methods of assessing evidence against theoretical predictions.

Submitted 10 May 2022
Accepted 6 February 2023
Published 9 March 2023

Corresponding authors
Benjamin G. Farrar, farrarbg@gmail.com
Ljerka Ostojić, lj.ostojic@uniri.hr

Academic editor
Diogo Provete

Additional Information and Declarations can be found on page 16

DOI **10.7717/peerj.14963**

Copyright
2023 Farrar et al.

Distributed under
Creative Commons CC-BY 4.0

OPEN ACCESS

How to cite this article Farrar BG, Vernouillet A, Garcia-Pelegrin E, Legg EW, Brecht KF, Lambert PJ, Elsherif M, Francis S, O'Neill L, Clayton NS, Ostojić L. 2023. Reporting and interpreting non-significant results in animal cognition research. *PeerJ* **11**:e14963 http://doi.org/10.7717/peer

Subjects Animal Behavior, Zoology, Statistics
Keywords Animal behavior, Animal cognition, Null hypothesis significance testing, Non-significant results, Statistical inferences, Negative results

INTRODUCTION

Null hypothesis significance testing (NHST) is a primary method of statistical analysis in animal cognition research. However, when NHST produces results that are not statistically significant, these are often difficult to interpret. If researchers test null hypotheses of zero effect (*i.e.,* there are no differences between groups or conditions), a non-significant result could result from a lack of any effect in the population (a true negative), or a failure to detect some true difference (a false negative). While current guidance encourages researchers to design studies with high statistical power to detect theoretically interesting effect sizes (*Lakens, 2017*; *Lakens, 2021*)—which can provide context for negative results—power analyses appear infrequent (*Fritz, Scherndl & Kühberger, 2013*). Hence, how statistically non-significant results are reported and interpreted within the NHST framework has been criticized on several grounds (*Fiedler, Kutzner & Krueger, 2012*; *Gigerenzer, Krauss & Vitouch, 2004*; *Lambdin, 2012*; *Vadillo, Konstantinidis & Shanks, 2016*). The most prominent criticism is that researchers often misreport a non-significant difference between test conditions or groups as if the result in each of the conditions or groups was exactly the same, and/or researchers misinterpret a non-significant results as evidence of the absence of an effect in regard to a substantive claim (*Aczel et al., 2018*; *Fidler et al., 2006*; *Hoekstra et al., 2006*). This misreporting or misinterpretation may even occur when the null hypothesis being considered is very likely to be incorrect (*Cohen, 1994*; *Gelman & Carlin, 2014*). Given these concerns, this study explored how animal cognition researchers report and interpret statistically non-significant results using a manually extracted dataset of negative claims following NHST from over 200 articles.

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 be *not* 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 researchers' data or more extreme data if the null hypothesis and all its assumptions were true, $\Pr(d(X) \geq d(x_0); H_0)$. If the *p*-value is lower than a pre-specified threshold (the α level), the statistical null hypothesis (H_0) is rejected in favor of an alternative hypothesis (*Neyman & Pearson, 1933*), whereas if the *p*-value is larger than the pre-specified threshold, H_0 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 (*Lambdin, 2012*). Formally, 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, Nelson & Simonsohn, 2016*), they can say that in the long run they would not reject H_0 more than 5% of the time, if H_0 were true. Similarly, if the design of the study is such that the statistical test had 90% power to detect the smallest effect size of interest, in the long run the researchers would only fail to reject H_0 10% of the time, if the smallest effect size of interest did exist in the population.

Without performing further analyses, it can be an error to conclude that there is evidence in favor of the null hypothesis following a non-significant result. The arbitrary nature of the α level highlights this: as an example, let us assume that 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. However, if we had chosen an α level of .10 instead, we would 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 and alternative hypotheses (*Lakens et al., 2018*). 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 or conditions or evidence of no effect when this inference is not necessarily warranted (*Aczel et al., 2018*; *Fidler et al., 2006*; *Hoekstra et al., 2006*). 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 and statistical test combinations with high power of statistical tests to detect the minimum effect size of theoretical interest (*Farrar, Boeckle & Clayton, 2020*).

While 'accepting the null' may be an error, just how severe an error it is requires discussing. 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 article, Neyman and Pearson often discussed 'accepting H_0' 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 ($p > 0.05$)" with "there was no statistically significant difference between the conditions", then the "serious mistake" of accepting the null becomes an issue of precision in language, rather than an egregious error.

The aim of this study was to explore how authors in fields related to animal cognition report and interpret statistically non-significant results by building on the methods used in similar studies in psychology and conservation biology (*Aczel et al., 2018*; *Fidler et al., 2006*; *Hoekstra et al., 2006*). This is an important step towards (i) identifying how often conclusions in animal cognition might be the result of NHST misreporting or misinterpretation, and (ii) highlighting areas in which animal cognition researchers can improve their statistical inferences and statistical reporting.

Table 1 Sources of articles containing negative results in their abstracts.

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

MATERIALS & METHODS

Data extraction and classification

We manually extracted data from a total of 20 sources, comprising 18 peer-reviewed journals in animal cognition, behavior, and welfare, one pre-print server, and articles recommended through Peer Community in Animal Science. The 20 sources are detailed in Table 1. Sources were screened backwards from March 2021, until 15 articles were identified that contained negative statements in titles or abstracts that corresponded to statistically 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 articles that they were confident with, explaining why three journals had 16 or 17 articles extracted.

Nine of the authors (author acronyms: BGF, AV, KB, EGP, LoN, PL, SF, EL, and ME) performed the coding and were each assigned two journals, except BGF who conducted the coding for four journals. Each coder screened the abstracts of each article of their assigned journals and identified any statements that was either, (i) reporting the results of a statistically non-significant hypothesis test (what we referred to as 'sample claim' in the Coding guidelines, https://osf.io/84puf/, as *Aczel et al., 2018*) and/or, (ii) interpreted a statistically non-significant result in relation to a substantive claim (what we referred to as 'population claim' in the Coding guidelines, https://osf.io/84puf/, as *Aczel et al., 2018*). If at

least one of these statements was present, the coder recorded the article's information (title, first author, journal, and year) and the statement(s) in question. For articles with multiple negative statements within each of the categories (reporting of results, interpretation of results in relation to a substantive claim), the coder recorded the statement that they thought was most clearly related to the paper's main claim, such that for each article, we had a maximum of one result statement and one substantive interpretation statement. Next, the coder verified that the statements were based on results from 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, the coder extracted the test result that they thought was most relevant to the claim. If the abstract claim was equally supported by multiple studies or experiments, the coder extracted the information from the first study or experiment presented.

After the title, abstract claim(s) (reporting of results and interpretations of the results in relation to a substantive claim), result text and *p*-value had been extracted, the coder categorized how each statement was phrased. Through piloting, discussion and from looking at previous studies (*Aczel et al., 2018*; *Fidler et al., 2006*; *Hoekstra et al., 2006*), we developed three categories. For the statements that reported the results in the title and abstract and for the result text, these were: (1) "Formally Correct, *i.e.,* Non-Significant" statements that either there was no *significant* difference between testing groups or conditions, or words to that effect, or a correct directional statement (text slightly altered from original phrasing; for the original phrasing see our Coding Guidelines, https://osf.io/84puf/); (2) "No Effect" statements that there was not a difference between testing groups or conditions, when in fact there was—it was just not significant in the analysis (text slightly altered from original phrasing; for the original phrasing see our Coding Guidelines, https://osf.io/84puf/); (3) "Ambiguous, Similar or Small Effect Size" statements about the results that neither suggest that the testing groups or conditions were the same, nor that there was no significant difference between them (which were later split into "Ambiguous" and "Similar or Small Effect Size" categories; text slightly altered from original phrasing; for the original phrasing see our Coding Guidelines, https://osf.io/84puf/). In addition to these descriptions, we developed a table of hypothetical statements that are detailed in Table 2, which were available to the coders during the project.

Similarly, the title, if it contained a statement referring to a statistically non-significant result and interpretations of the results in relation to a substantive claim from the abstracts were categorized into three categories: (1) "Correct, *i.e.,* Justified": An interpretation 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 regards to the substantive claim, or that the study provides no strong evidence of a difference (text slightly altered from original phrasing; for the original phrasing see our Coding Guidelines, https://osf.io/84puf/), (2) "Caveated or Ambiguous": An interpretation of the non-significant results as suggesting/indicating etc. that X and Y do not differ, or showing that they are similar, and (3) "No Effect": An incorrect interpretation of the non-significant result as showing that X and Y do not differ in relation to a substantive

Table 2 Example categorization of sample-level statements.

Category	Non-Significant	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.	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 and Y were similar. There was no large/clear difference between X and Y. There was no large effect of X on Y.

claim (text slightly altered from original phrasing; for the original phrasing see our Coding Guidelines, https://osf.io/84puf/). In addition to these descriptions, we developed a table of hypothetical statements that are detailed in Table 3.

Reliability and quality control

Twenty-four articles (8.5%) were double-blind coded 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 in the extracted dataset before the data were used in any analysis.

Double-blind extraction

To test the reliability of the coding scheme, 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, we computed inter-rate agreement for each variable that we extracted (title statement, reporting of results in abstract, interpretation of results in relation to a substantive claim in abstract, reporting of results in result section, *p*-value), including raw percentage agreement, Cohen's kappa and Gwet's AC1.

Quality control

To minimize mistakes in the dataset, all articles underwent the quality control procedure. Here, a second coder reviewed the data extracted from each article. Ten of the authors (BGF, AV, KB, EGP, LoN, PL, SF, EL, ME, and LO) served as second coders, and each was assigned one other coder's original set of articles to quality control. The quality controller verified (1) that a claim from the title/abstract has been extracted, (2) that any

Table 3 Example categorization 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.	…suggesting that X is not related to Y.	…meaning that X is not related to Y.
		…indicating that X is not related to Y.	…showing that X is not related to Y.
		…suggesting/indicating that there is no difference between X and Y.	There is no difference between X and Y.
			X and Y do not differ.
		…suggesting that X has not changed Y.	X and Y are similar.
	In addition to being not statistically different to each 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.		X and Y are the same (show the same effect, etc.).
		Our results provide no strong evidence that X and Y are different.	X does not change Y.
		…suggesting that X and Y are similar.	Our results provide no evidence that X and Y are different.

claim extracted referred to a statistically non-significant result, (3) that the result that was extracted corresponded to the claim that was extracted, and (4) that 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.

Analysis

We present the percentage of claims in each category across the extracted variables (title statements, statements reporting non-significant results in the abstract, statements interpreting the results in relation to a substantive claim in abstracts, and reporting of the statistical result in the results section). To illustrate the types of claims placed in each category, examples that we felt were particularly representative of each are provided in tables. In addition, every classification can also be viewed in the open dataset at https://osf.io/84puf/. We used a Chi-squared test to test whether, if the results reported "No Effect", it was more likely that a "No Effect" interpretation would also be made in the abstract than when the results were correctly reported as "Non-Significant".

In addition to our primary descriptive analysis, we performed an exploratory analysis of the *p*-value distribution of the *p*-values associated with all extracted non-significant results. We used a two-sided Kolmogorov–Smirnov test to compare the observed distribution to

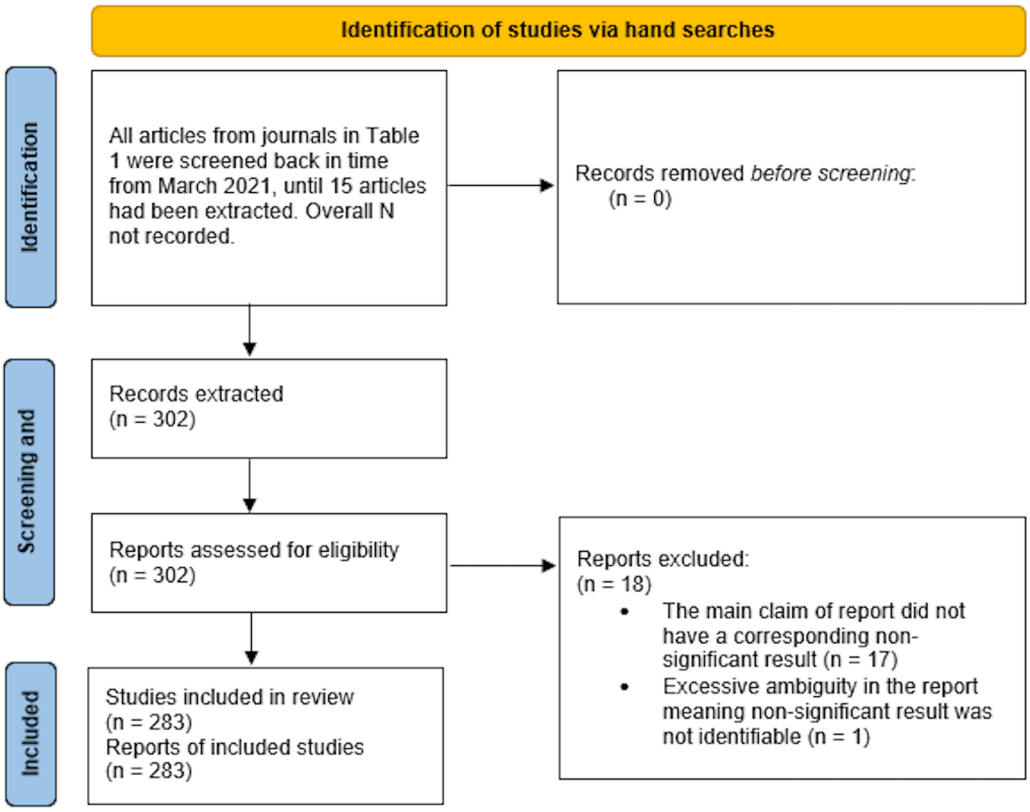

Figure 1 **PRISMA figure.**

Full-size DOI: 10.7717/peerj.14963/fig-1

a uniform distribution, *i.e.,* the theoretical distribution that the *p*-values would have been samples from if the null hypothesis of no difference was true for all studies.

RESULTS

We extracted data from 302 articles. Of these, 18 were excluded due to their identified claim having no corresponding negative result of NHST (*e.g.*, only descriptive statistics 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 (Fig. 1). The results of the Reliability and Quality Control coding can be found in the Supplemental Material.

Title statements

Forty-four titles (19% of the total articles) were identified as containing statements resulting from non-significant results of NHSTs. Of these, 37 (84%) were classified as interpreting the non-significant result as evidence of no effect, whereas seven (16%) were classified as caveated claims or claims about testing groups or conditions being 'similar'. Table 4 provides examples of these claims.

Table 4 Examples of claims in the titles of articles following non-significant NHST classified as "No Effect" and "Caveated or Similar".

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)

Abstract statements

Reporting of results in abstracts

We extracted 278 claims that reported non-significant results of NHST. Of these, 174 (63%) were classified as claiming evidence of no effect, 71 (26%) as making formally correct statements that there were no statistically significant differences between groups or conditions, 17 (6%) as making claims about an effect being 'similar' between groups or conditions, or as describing a small effect size, and 16 (6%) were classified as ambiguous. Table 5 provides examples of these claims.

Interpretations of results in abstracts

We extracted 63 statements that were interpretations of statistically non-significant results in relation to substantive claims. Of these, 45 (71%) were classified as caveated and 18 as claiming that there was no effect (29%). Table 6 provides examples of these claims.

Result text

In the results sections, 276 non-significant results of NHST were coded. Of these, 140 (52%) were classified as reporting the results as "Non-Significant", 113 (41%) as reporting that there was "No Effect", 12 (4%) as reporting groups or conditions being "Similar", 10 (4%) were classified as "Ambiguous", and one (0.4%) as reporting a "trend" in the opposite direction to the prediction. Table 7 provides examples of the different types of result reporting.

Table 5 **Examples of claims about the sample in the abstracts of articles following non-significant NHST classified as "No Effect", "Similar or Small Effect Size", "Non-Significant" or "Ambiguous".**

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, Winters & Jawor (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, Clary & Kelly (2021)
"No significant correlations between degree of laterality and behavioral interest in the stimuli were found"
Lilley, De Vere & Yeater (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, Elias & Rosenthal (2021)
"No systematic variations on space allocation were observed in neither experiment"
Ribes-Iñesta, Hernández & Serrano (2020)

Notably, if a sentence reporting the results in the results section was classified as "No Effect", it was more likely that this statistical test would also be reported as "No Effect" in the abstract, compared to when the result was classified as "Non-Significant" ($\chi 2(1, N = 211) = 21.65$, $p < .0001$). Limiting the data to just those with responses in the abstract and results classified as "Non-Significant" or "No Effect", of the 92 statements in the results classified as "No Effect", 80 (87%) of the corresponding statements reporting the results in the abstract were classified as "No Effect". In contrast, of the 119 statements in the results classified as "Non-Significant", only 67 (56%) were reported as "No Effect" in the abstract. Nevertheless, the "No Effect" phrasing when reporting results in the abstracts was absolutely the most likely classifications for both "No Effect" and "Non-Significant" phrasings in the results section.

p-value distributions

In total, 202 of the 283 articles 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

Table 6 Examples of claims about populations in the abstracts of articles following non-significant NHST classified as "No Effect" and "Caveated, Ambiguous or Similar".

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, Booth & Wong (2021)
"…suggesting that size is not a determinant factor for feral horse society."
Pinto & Hirata (2020)

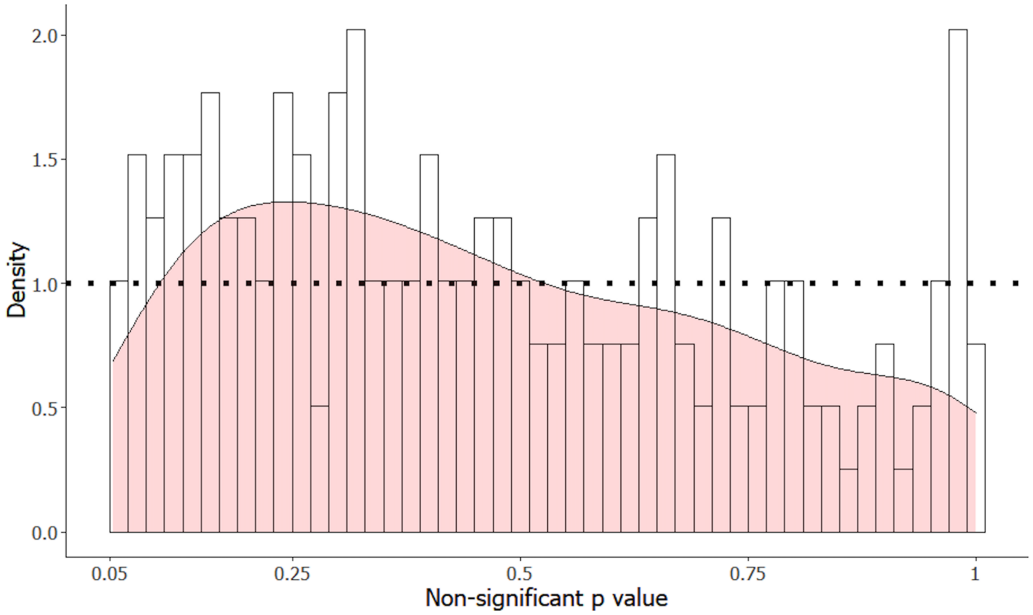

Figure 2 Distribution of non-significant *p*-values from result sections of 198 articles in animal cognition and related fields, with a density distribution overlaid in pink. The dotted line shows the average density.

Full-size DOI: 10.7717/peerj.14963/fig-2

.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 Fig. 2. This distribution significantly differs from a uniform distribution (two-sided Kolmogorov–Smirnov test, $D = 0.12$, $p = .0087$).

Table 7 Examples of statement reporting the results in the results sections of articles using non-significant NHST classified as "No Effect", "Similar or Small Effect Size", "Non-Significant" or "Ambiguous".

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, Broschard & Wasserman (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, Müller & Nieder (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 andNR 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)

Figure 3 contrasts the distribution of Fig. 2 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% statistical power to detect the true effect size of H1 if 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.

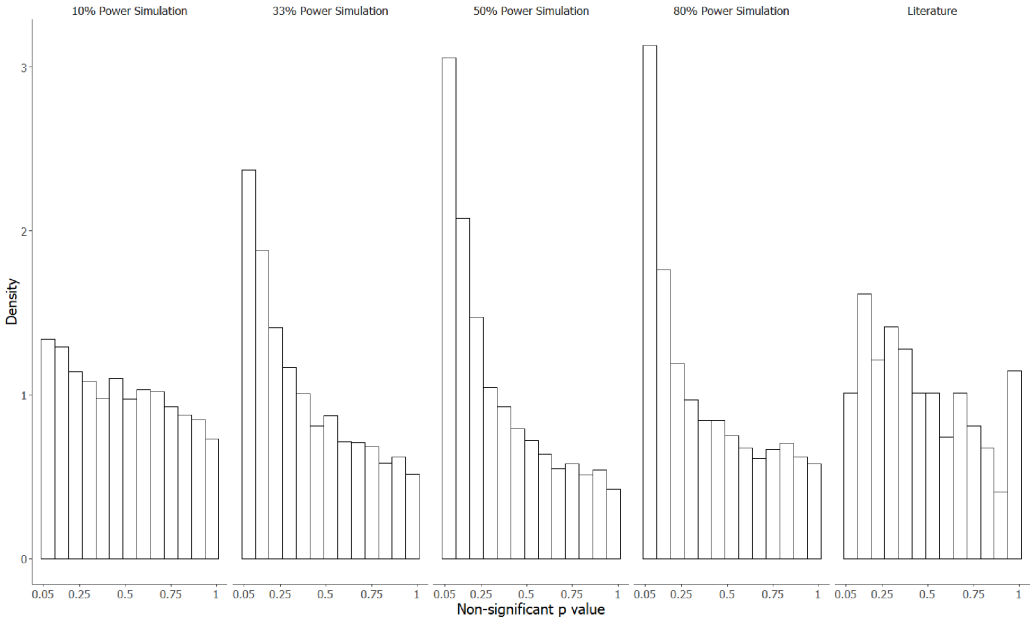

Figure 3 The observed *p*-value distribution of 198 *p*-values > .05, compared to three simulated distributions where 80% of alternative hypotheses were correct. The observed *p*-value distribution was manually extracted from results corresponding to negative claims present in the abstracts of animal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

Full-size DOI: 10.7717/peerj.14963/fig-3

DISCUSSION

We extracted and classified how animal cognition researchers reported and interpreted the results of non-significant null hypothesis significance tests in 253 articles between 2019 and 2021. Across titles, abstracts, and results, we classified non-significant results as being reported with the "No Effect" phrasing that has often been labelled as erroneous in 84% of titles with a statistically non-significant result, in 63% of abstracts reporting a statistically non-significant result and in 41% of results sections reporting a statistically non-significant result. Reporting statistically non-significant results as "Non-Significant" was less common in titles and abstracts, but as prevalent as "No Effect" phrasings in the result sections (titles: 16%; reporting of results in abstract: 26%; result text: 52%). The other, albeit less frequently classified method of reporting statistically non-significant results was to comment on the similarity between groups or conditions (reporting of results in abstracts: 6%; result text: 4%).

Overall, these results demonstrate considerable heterogeneity in how animal cognition researchers report and interpret non-significant results in published articles. However, we often found it difficult to categorize results due to the heterogeneity in how statements referring to statistically non-significant results were phrased. Despite this heterogeneity, our results suggest that statistically non-significant results are at risk of being misreported and misinterpreted 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 analyzing how these studies are cited, in other publications but also in media reports and student essays.

A good example for the different ways in which the same phrasing can be interpreted by different researchers within the same research community are the three instances in which authors phrased the results of an ANOVA term as "no main effect", which we classified as "Ambiguous" as per our Coding guidelines. However, during the review process, one of the reviewers stated that they treat "no main effect" as formally equivalent to "no effect". The reason for our original classification was that referring to "no main effect" gives the reader more information about the statistical analysis used and thus may be more likely to be interpreted by readers correctly as the analysis yielding a non-significant main effect compared as when "no effect" is used. However, the reviewer's interpretation is as justifiable, and this example clearly illustrates the importance of investigating how researchers interpret and cite original findings in their own publications.

Possibly encouragingly, when researchers extended "No Effect" statements from reporting their study's results to interpreting them in relation to a substantive claim, they routinely opted for qualifiers to caveat inference to the populations (*e.g.*, "…these results *suggest* that there is no effect at the population level"). However, such qualified statements lack precision and are open multiple interpretations—they make only a vague suggestion that the strength of the evidence for the claim might be low, a claim that can often be explored in a quantitative and precise manner. Moreover, it is likely that such caveating is not unique to statistically non-significant results but also used to caveat significant findings, too. If that is correct, then the caveating may have more to do with researchers being critical of, or attempting to appear critical, of their results in general, and acknowledging that there may be alternative conceptual interpretations of their results (*Farrar & Ostojić, 2019*), rather than being specific to recognising the lack of information associated with studies with low power of statistical tests. However, more research is needed to pinpoint *precisely* how such statements are interpreted and implemented by scientists and the wider community. As already noted, one way in which researchers might reduce the ambiguity of their negative statements would be to use more formal methods of assessing evidence against informative null hypotheses, such as by testing against theoretically interesting effect sizes using as equivalence tests or comparing plausible null and alternative hypotheses using Bayes factors. Although beyond the scope of the current project, *Lakens (2017)* provides a detailed tutorial for equivalence testing in psychological research, and *Rose et al. (2018)* in animal behavior, and *Rouder et al. (2009)* provide an introduction to Bayes Factors. In addition, we would like to refer the reader to a competence model developed by *Edelsbrunner & Thurn (2020)* for all researchers who are involved in teaching statistics and mentoring students in the field of animal behavior science.

Notably, our coding team found it difficult to identify whether interpretations of a study's results in relation to a substantive claim in the abstracts were based on a statistically non-significant result and thus also to classify them in the next step. This difficulty 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 decision between a null and alternative hypothesis in a null hypothesis significance test.

Finally, we classified the formally incorrect reporting of results and interpretations of results as "No Effect" more commonly in abstracts and titles than "No Effect" reporting of results in the results section. That is, authors who have written out "Non-Significant" results in the results section nevertheless used the "No Effect" phrasing for reporting and interpreting the results in the abstracts and titles. This could be due to two factors, namely word limits and incentives to make bolder claims. If this is correct, then the former should be considered by journal editorial boards when setting their policy.

The *p*-value distribution likely differed from a uniform distribution for two reasons: the cumulative frequency was greater in the observed distribution for smaller *p*-values ($p < .3$) and was also greater for large *p*-values ($p > .95$). The larger density of smaller *p*-values is consistent with research with low-powered statistical tests in which the null hypothesis was incorrect, but which produces *p*-values that did not reach statistical significance. The density of very large *p*-values is consistent with researchers applying corrections that might increase *p*-values, such as Bonferroni corrections, or by using statistical tests with small sample sizes that produce non-uniform *p*-value distributions under the null hypothesis. An interesting contrast between the observed and simulated *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 simulated distributions. This is likely because we extracted results that we as coders had interpreted as being statistically non-significant for the manual dataset, but *p*-values in the range .05−0.1 are often interpreted by the original authors as "trends" or "marginally significant" and may therefore lead to author interpretations as if there had not been a statistically non-significant result.

CONCLUSIONS

This study explored reporting and interpretation of statistically non-significant results in animal cognition literature through classification by other researchers in the field. In line with previous studies in other disciplines (*Aczel et al., 2018*; *Fidler et al., 2006*), we found that statistically non-significant results were often reported as if there were no differences observed between groups or conditions, and this was the case in the titles, abstracts and result sections of papers, although it was most frequent in the titles and abstracts. These results suggest that incorrect theoretical inferences based on non-significant results in animal cognition literature are common. However, 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 but may be grave. Nevertheless, these findings 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 formally correct, and (ii) use more formal methods of assessing the evidence against theoretical predictions.

ACKNOWLEDGEMENTS

We would like to thank Balazs Aczel for discussions and clarifications about previous research in this area

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

Benjamin G. Farrar was supported by the University of Cambridge BBSRC Doctoral Training Programme (BB/M011194/1). Alizée Vernouillet is currently supported by a BOF fellowship (BOF.PDO.2021.0035.01). Katharina F. Brecht was supported by a DFG Grant (BR 5908/1-1) and by a University of Tübingen Athene Fellowship. Mahmoud Elsherif is currently supported by The Baily Thomas Charitable Fund (TRUST/VC/AC/SG/5843-8995). Edward W. Legg is supported by a MSCA Fellowship (INTOM-794270). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
University of Cambridge BBSRC Doctoral Training Programme: BB/M011194/1.
BOF fellowship: BOF.PDO.2021.0035.01.
DFG Grant: BR 5908/1-1.
University of Tübingen Athene Fellowship.
The Baily Thomas Charitable Fund: TRUST/VC/AC/SG/5843-8995.
MSCA Fellowship: INTOM-794270.

Competing Interests

Ljerka Ostojić is an Academic Editor for PeerJ.

Author Contributions

- Benjamin G. Farrar conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Alizée Vernouillet performed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Elias Garcia-Pelegrin performed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Edward W. Legg performed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Katharina F. Brecht performed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Poppy J. Lambert performed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Mahmoud Elsherif performed the experiments, authored or reviewed drafts of the article, and approved the final draft.

- Shannon Francis performed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Laurie O'Neill performed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Nicola S. Clayton conceived and designed the experiments, authored or reviewed drafts of the article, and approved the final draft.
- Ljerka Ostojić conceived and designed the experiments, performed the experiments, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The data are available at OSF: Farrar, Benjamin G. 2022. "Non-Significant Results in Animal Cognition." OSF. October 30. osf.io/gdp6f.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.14963#supplemental-information.

REFERENCES

Aczel B, Palfi B, Szollosi A, Kovacs M, Szaszi B, Szecsi P, Zrubka M, Gronau QF, 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(3)**:357–366 DOI 10.1177/2515245918773742.

Anselme P, Robinson MJF. 2019. Evidence for motivational enhancement of sign-tracking behavior under reward uncertainty. *Journal of Experimental Psychology: Animal Learning and Cognition* **45(3)**:350–355 DOI 10.1037/xan0000213.

Aparecida Martins R, Ribeiro Caldara F, Crone C, Markiy Odakura A, Bevilacqua A, Oliveira dos Santos Nieto VM, Aparecida Felix G, Pereira dos Santos A, Sousa dos Santos L, Garófallo Garcia R, De Castro Lippi IC. 2021. Strategic use of straw as environmental enrichment for prepartum sows in farrowing crates. *Applied Animal Behaviour Science* **234**:105194 DOI 10.1016/j.applanim.2020.105194.

Beran MJ, French K, Smith TR, Parrish AE. 2019. Limited evidence of number–space mapping in rhesus monkeys (*Macaca mulatta*) and capuchin monkeys (Sapajus apella). *Journal of Comparative Psychology* **133(3)**:281–293 DOI 10.1037/com0000177.

Brecht KF, Müller J, Nieder A. 2020. Carrion crows (*Corvus corone corone*) fail the mirror mark test yet again. *Journal of Comparative Psychology* **134(4)**:372–378 DOI 10.1037/com0000231.

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(1)**:107–119 DOI 10.1007/s10071-020-01425-9.

Cohen J. 1994. The Earth is round ($p < .05$). *American Psychologist* **49(12)**:997–1003 DOI 10.1037/0003-066X.49.12.997.

Cunningham PJ, Shahan TA. 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 DOI 10.1037/xan0000245.

DeVries MS, Winters CP, Jawor JM. 2020. Similarities in expression of territorial aggression in breeding pairs of northern cardinals, Cardinalis cardinalis. *Journal of Ethology* **38(3)**:377–382 DOI 10.1007/s10164-020-00659-x.

Edelsbrunner PA, Thurn C. 2020. Improving the Utility of Non-Significant Results for Educational Research. *PsyArXiv* DOI 10.31234/osf.io/j93a2.

Farrar BG, Boeckle M, Clayton NS. 2020. Replications in comparative cognition: What should we expect and how can we improve? *Animal Behavior and Cognition* **7(1)**:1–22 DOI 10.26451/abc.07.01.02.2020.

Farrar BG, Ostojić L. 2019. The Illusion of Science in Comparative Cognition. *PsyArXiv* DOI 10.31234/osf.io/hduyx.

Fidler F, Burgman MA, 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(5)**:1539–1544 DOI 10.1111/j.1523-1739.2006.00525.x.

Fiedler K, Kutzner F, Krueger JI. 2012. The long way from α-error control to validity proper: problems with a short-sighted false-positive debate. *Perspectives on Psychological Science* **7(6)**:661–669 DOI 10.1177/1745691612462587.

Fritz A, Scherndl T, Kühberger A. 2013. A comprehensive review of reporting practices in psychological journals: are effect sizes really enough? *Theory & Psychology* **23(1)**:98–122 DOI 10.1177/0959354312436870.

Gelman A, Carlin J. 2014. Beyond power calculations. *Perspectives on Psychological Science* **9(6)**:641–651 DOI 10.1177/1745691614551642.

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. In: *Handbook on Quantitative Methods in the Social Sciences*. Thousand Oaks: Sage, 389–406.

Goodman S. 2008. A dirty dozen: twelve *p*-value misconceptions. *Seminars in Hematology* **45(3)**:135–140 DOI 10.1053/j.seminhematol.2008.04.003.

Guadarrama SS, Domínguez-Vega H, Díaz-Albiter HM, 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(1)**:74–82 DOI 10.1080/10888705.2018.1562350.

Harris JA, Bouton ME. 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(3)**:256–272 DOI 10.1037/xan0000242.

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 DOI 10.19227/jzar.v8i4.523.

Hoekstra R, Finch S, Kiers HAL, Johnson A. 2006. Probability as certainty: dichotomous thinking and the misuse of *p* values. *Psychonomic Bulletin & Review* **13(6)**:1033–1037 DOI 10.3758/BF03213921.

Kawaguchi M, Kuriwada T. 2020. Effect of predator cue on escape and oviposition behaviour of freshwater snail. *Behaviour* **157(7)**:683–697 DOI 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(4)**:488–495 DOI 10.1037/com0000182.

Koczura M, Martin B, Musci M, Massimo MD, 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 DOI 10.3389/fvets.2020.612504.

Kvarnemo C, Andersson SE, Elisson J, Moore GI, Jones AG. 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(2)**:235–248 DOI 10.1007/s10164-021-00698-y.

Lakens D. 2017. Equivalence tests: a practical primer for *t* tests, correlations, and meta-analyses. *Social Psychological and Personality Science* **8(4)**:355–362 DOI 10.1177/1948550617697177.

Lakens D. 2021. Sample Size Justification. DOI 10.31234/osf.io/9d3.

Lakens D, Adolfi FG, Albers CJ, Anvari F, Apps MAJ, Argamon SE, Baguley T, Becker RB, Benning SD, Bradford DE, Buchanan EM, Caldwell AR, Van Calster B, Carlsson R, Chen S-C, Chung B, Colling LJ, Collins GS, Crook Z, Cross ES, Daniels S, Danielsson H, DeBruine L, Dunleavy DJ, Earp BD, Feist MI, Ferrell JD, Field JG, Fox NW, Friesen A, Gomes C, Gonzalez-Marquez M, Grange JA, Grieve AP, Guggenberger R, Grist J, Van Harmelen A-L, Hasselman F, Hochard KD, Hoffarth MR, Holmes NP, Ingre M, Isager PM, Isotalus HK, Johansson C, Juszczyk K, Kenny DA, Khalil AA, Konat B, Lao J, Larsen EG, Lodder GMA, Lukavský J, Madan CR, Manheim D, Martin SR, Martin AE, Mayo DG, McCarthy RJ, McConway K, McFarland C, Nio AQX, Nilsonne G, Linode Oliveira C, Orban de Xivry J-J, Parsons S, Pfuhl G, Quinn KA, Sakon JJ, Saribay SA, Schneider IK, Selvaraju M, Sjoerds Z, Smith SG, Smits T, Spies JR, Sreekumar V, Steltenpohl CN, Stenhouse N, Świątkowski W, Vadillo MA, Van Assen MALM, Williams MN, Williams SE, Williams DR, Yarkoni T, Ziano I, Zwaan RA. 2018. Justify your alpha. *Nature Human Behaviour* **2(3)**:168–171 DOI 10.1038/s41562-018-0311-x.

Lambdin C. 2012. Significance tests as sorcery: science is empirical—significance tests are not. *Theory & Psychology* **22(1)**:67–90 DOI 10.1177/0959354311429854.

Lazarowski L, Thompkins A, Krichbaum S, Waggoner LP, Deshpande G, Katz JS. 2020. Comparing pet and detection dogs (*Canis familiaris*) on two aspects of social cognition. *Learning & Behavior* **48(4)**:432–443 DOI 10.3758/s13420-020-00431-8.

Lilley MK, De Vere AJ, Yeater DB. 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**:1–11 DOI 10.46867/ijcp.2020.33.03.01.

Mayo DG. 2018. *Statistical inference as severe testing: how to get beyond the statistics wars*. Cambridge: Cambridge University Press.

Meza P, Elias DO, Rosenthal MF. 2021. The effect of substrate on prey capture does not match natural substrate use in a wolf spider. *Animal Behaviour* **176**:17–21 DOI 10.1016/j.anbehav.2021.03.014.

Neyman J. 1976. Tests of statistical hypotheses and their use in studies of natural phenomena. *Communications in Statistics—Theory and Methods* **5(8)**:737–751 DOI 10.1080/03610927608827392.

Neyman J, Pearson ES. 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(694–706)**:289–337 DOI 10.1098/rsta.1933.0009.

O'Donoghue EM, Broschard MB, Wasserman EA. 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(2)**:107–123 DOI 10.1037/xan0000234.

Paijmans KC, Booth DJ, Wong MYL. 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(2)**:125–134 DOI 10.1111/eth.13110.

Pereira FC, Teixeira DL, Boyle LA, Pinheiro Machado Filho LC, Williams SRO, 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 DOI 10.3389/fvets.2020.620810.

Piefke TJ, Bonnell TR, De Oliveira GM, Border SE, Dijkstra PD. 2021. Social network stability is impacted by removing a dominant male in replicate dominance hierarchies of a cichlid fish. *Animal Behaviour* **175**:7–20 DOI 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 DOI 10.1016/j.beproc.2020.104277.

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 DOI 10.1016/j.beproc.2020.104256.

Rose EM, Mathew T, Coss DA, Lohr B, Omland KE. 2018. A new statistical method to test equivalence: an application in male and female eastern bluebird song. *Animal Behaviour* **145**:77–85 DOI 10.1016/j.anbehav.2018.09.004.

Rouder JN, Speckman PL, Sun D, Morey RD, Iverson G. 2009. Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review* **16**(**2**):225–237 DOI 10.3758/PBR.16.2.225.

Simmons JP, Nelson LD, Simonsohn U. 2016. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. In: Kazdin AE, ed. *Methodological issues and strategies in clinical research*. American Psychological Association, 547–555 DOI 10.1037/14805-033.

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 DOI 10.1016/j.beproc.2020.104290.

Stevens A, Doneley R, Cogny A, Phillips CJC. 2021. The effects of environmental enrichment on the behaviour of cockatiels (*Nymphicus hollandicus*) in aviaries. *Applied Animal Behaviour Science* **235**:105154 DOI 10.1016/j.applanim.2020.105154.

Vadillo MA, Konstantinidis E, Shanks DR. 2016. Underpowered samples, false negatives, and unconscious learning. *Psychonomic Bulletin & Review* **23**(**1**):87–102 DOI 10.3758/s13423-015-0892-6.

Vernouillet A, Clary D, Kelly DM. 2021. Highly social pinyon jays, but not less social Clark's nutcrackers, modify their food-storing behaviour when observed by a heterospecific. *BioRxiv* DOI 10.1101/2021.02.28.433225.

Wu Y, Petrosky AL, Hazzi NA, Woodward RL, 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 DOI 10.1016/j.anbehav.2021.02.015.

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 DOI 10.1016/j.applanim.2020.105182.