Reporting and interpreting non-significant results in animal cognition research

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

[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, Faculty of Humanities and Social Sciences, University of Rijeka, RIjeka, Croatia

[5] Division of Cognitive Sciences, University of Rijeka, Rijeka, Croatia

[6] Centre for Mind and Behaviour, University of Rijeka, Rijeka, Croatia

[7] Institute for Neurobiology, University of Tuebingen, Tuebingen, Germany

[8] Messerli Research Insititute, University of Vienna, Vienna, Austria

[9] Department of Psychology, University of Birmingham, Birmingham, United Kingdom

[10] Comparative Cognition Research Group, Max Planck Institute for Ornithology, Seewiesen, Germany

Corresponding Authors: Benjamin G Farrar, Ljerka Ostojic
Email address: farrarbg@gmail.com, lj.ostojic@uniri.hr

How negative results are reported and interpreted following null hypothesis significance testing is often criticized. With small sample sizes and often low number of test trials, studies in animal cognition are prone to producing non-significant *p*-values, irrespective of whether this is a false negative or true negative result. Thus, we assessed how negative results are reported and interpreted across published articles in animal cognition and related fields. In this study, 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. We found a large amount of heterogeneity in how researchers report non-significant *p*-values in the result sections of articles, and how they interpret them in the titles and abstracts. “No Effect” interpretations were common in the titles (84%), abstracts (64%), and results sections (41%) of papers, whereas “Non-Significant” interpretations were less common in the titles (0%) and abstracts (26%), but were 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 research to detect effect sizes of interest.

Reporting and interpreting non-significant results in animal cognition research

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

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

Corresponding Authors:
Benjamin G. Farrar
20 Bosworth Road
Cambridge
CB1 8RG
United Kingdom
Email address: farrarbg@gmail.com

Ljerka Ostojić
Odsjek za psihologiju
Filozofski fakultet u Rijeci
Sveučilišna avenija 4
51000 Rijeka
Croatia
Email address: lj.ostojic@uniri.hr

Abstract

How negative results are reported and interpreted following null hypothesis significance testing is often criticized. With small sample sizes and often low number of test trials, studies in animal cognition are prone to producing non-significant *p*-values, irrespective of whether this is a false

negative or true negative result. Thus, we assessed how negative results are reported and interpreted across published articles in animal cognition and related fields. ~~In this study,~~ 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. We found a large amount of heterogeneity in how researchers report non-significant *p*-values in the result sections of articles, and how they interpret them in the titles and abstracts. "No Effect" interpretations were common in the titles (84%), abstracts (64%), and results sections (41%) of papers, whereas "Non-Significant" interpretations were less common in the titles (0%) and abstracts (26%), but were 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 research to detect effect sizes of interest.

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, 2021) – which can provide context for negative results – power analyses appear infrequent (Fritz et al., 2013). Hence, how negative results are reported and interpreted following non-significant results from NHST has been criticised on several grounds (Gigerenzer et al., 2004; Lambdin, 2012), with some researchers noting that false negative claims can inhibit scientific progress just as much as false positive claims (Fiedler et al., 2012; Vadillo et al., 2016). The most prominent criticism is that researchers often misreport or misinterpret non-significant results as meaning that, i) within the **sample** that *was* tested there was no effect (a specific report about what happened in the study) and/or that ii) the non-significant result means that in the target **population** in general there *is* no effect (Aczel et al., 2018; Fidler et al., 2006; Hoekstra et al., 2006). This misreporting or misinterpretation even

occurs even 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 non-significant results using a manually extracted dataset of negative claims following NHST from over 200 articles.

1. NHST and *p*-values

When using NHST, researchers attempt to reject a statistical model (the null hypothesis) with their data while controlling the rate at which they will make false-positive decisions in the long-term (Neyman & Pearson, 1933). Most often, this statistical null is that there is absolutely no difference between two groups or conditions (for example a mean difference of 0 for a *t*-test; 'nil' hypothesis; Cohen, 1994), or, in the case of a one-tailed test, that the difference will not be zero *or* that it will 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 et al. 2011), 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.

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

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, when performing this procedure. 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 of conditions or evidence of no effect when this inference is not necessarily warranted. For example, Hoekstra et al., (2006) reported that 41% of articles containing non-significant results in 1994 and 1995 in *Psychonomic Bulletin & Review* included the interpreted of non-significant results as "evidence of no effect", a figure which rose to 60% in 2002 to 2004. Similarly, Fidler et al. (2006) found that 63% of articles in 2000 and 2001 in *Conservation Biology* and *Biological Conservation* reported non-significant findings as "evidence of no effect". More recently, Aczel et al. (2018) found that 72% of non-significant results were reported as "no effect" in the abstracts of 2015 articles in *Psychonomic Bulletin & Review*, *Journal of Experimental Psychology: General*, and *Psychological Science*. Such an error might be especially important in animal cognition research, in which a combination of small sample sizes and low trial number may limit the ability of researchers to design studies and statistical test combinations with high power of statistical tests to detect the minimum effect size of theoretical interest (Farrar et al., 2020).

While 'accepting the null' may be an error, just how severe an error it is requires discussing. In their survey of 86 *Psychonomic Bulletin & Review* Hoekstra et al. (2006, p. 1036) reported that: "We found the *serious* mistake of accepting the null hypothesis and claiming no effect in 60% (CI: 53%, 66%) of the articles that reported statistically nonsignificant results" (emphasis added). And interpreting a non-significant result as if there were no differences between conditions ranks at Number 2 of Goodman's (2008) "Dirty Dozen" *p*-value misconceptions. However, just because a researcher might report the results of significance tests incorrectly, this does not mean that they themselves, or their readers, necessarily interpreted the significance test incorrectly. In their 1933 paper, Neyman 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 ($p > 0.05$)" or "therefore we accept H_0" with "there was no statistically significant difference between the conditions" or "therefore we do no reject H_0", then the "serious mistake" of accepting the null becomes an issue of precision in language, rather than an egregious error. This is exemplified in cases where the observed experimental data are clearly more in-line with the null hypothesis than the alternative hypothesis, H_1.[1]

~~3. Exploring non-significant result reporting and interpretation in animal cognition~~

~~Understanding how animal cognition researchers have reported and interpreted non-significant findings in their published articles is an important step to, i) identify how often negative conclusions in animal cognition might be a result of NHST misreporting or misinterpretation, and ii) highlight areas in which animal cognition researchers can improve their statistical inferences and statistical reporting. In this study we explored how authors in fields related to animal cognition report and interpret 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). Furthermore, we also extracted the *p*-values associated with the negative results in our sample. We compared the distribution of these to four simulated *p*-value distributions in which research was performed with no publication bias and either 10%, 33%, 50%, or 80% power. Comparing the observed literature p-value distribution to the four simulated distribution can provide cues to the average statistical power of the research we extracted data from, as well as the presence of any biasing effects like publication bias (but note there is likely a large degree of heterogeneity here – see Nord et al. (2017) for a discussion of this in neuroscience research).~~

~~4. Reporting and interpreting non-significant results in animal cognition~~

[1] For example, consider a study in which birds' latencies to approach a novel object are compared between 10,000 wild and 10,000 hand reared birds (with 99% power to detect a pre-specified effect size of interest of 2 seconds), and in which a difference of 0.02 seconds was observed. This difference may even be statistically significant, but the minimum effect size of interest. Here, saying "there was no difference between the latency of wild and hand reared birds to approach the novel object ($p > .05$)", although literally incorrect, does not seem to be an error of great consequence.

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

Materials & Methods

Sample

We extracted data from a total of 20 sources, comprising 18 peer-reviewed journals, one pre-print server, and articles recommended through Peer Communities In. The 20 sources are detailed in *Table 1*.

196 **Data extraction and Classification**

197 BGF, AV, KB, EGP, LoN, PL, SF, EL, and ME performed the coding and were each
198 assigned two journals, except BGF who conducted the coding for four journals. Each coder
199 screened the abstracts of each article of their assigned journals and identified any negative
200 statements about either, i) the specific sample tested in that study or, ii) a wider population. If a
201 negative statement was present, the coder then recorded the paper's information (title, first
202 author, journal, and year) and the negative statement. For articles with multiple negative
203 statements for either the sample or the population, the coder recorded the negative statement that
204 they thought was most clearly related to the paper's main claim, such that for each article, we
205 had a maximum of one negative sample statement and one negative population statement. Next,
206 the coder verified that the statements were based on results from NHST. If verified, the coder
207 then extracted the text of the NHST that corresponded to the abstract claim from the results
208 section of the manuscript, including the associated *p*-value. If there was more than one
209 corresponding statistical test within an experiment, the coder extracted the test result that they
210 thought was most relevant to the claim. If the abstract claim was equally supported by multiple
211 studies or experiments, the coder extracted the information from the first study or experiment
212 presented.
213 After the title, abstract claims (sample and population), result text and *p*-value had been
214 extracted, the coder categorized how each negative statement was reported. Through piloting,
215 discussion, from looking at previous studies (Aczel et al., 2018; Fidler et al., 2006; Hoekstra et
216 al., 2006), we developed three categories. For the sample claims and result text, these were: 1)
217 "Non-Significant" statements that either there was no *significant* difference between two
218 conditions, or words to that effect, or a correct directional statement; 2) "No Effect" statements
219 that there was not a difference within the sample, when in fact there was — it was just not
220 significant in the analysis; 3) "Ambiguous, Similar or Small Effect Size" statements about the
221 results that neither suggest that samples were the same, nor that there was no significant
222 difference between them (which were later split into "Ambiguous" and "Similar or Small Effect
223 Size" categories). In addition to these descriptions, we developed a table of hypothetical
224 statements that are detailed in *Table 2,* which were available to the coders during the project.
225 Similarly, the title, if it contained a negative statement, and population claims from the abstracts
226 were categorized into three categories: 1) "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 the population, or that the study provides no strong evidence of a difference, 2) "Justified, Caveated or Ambiguous": An interpretation of the non-significant results as suggesting/indicating etc. that X and Y do not differ in the population, or showing that they are similar, and 3) "No Effect": An interpretation of the non-significant result as showing that X and Y do not differ in the population. In addition to these descriptions, 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 in order to assess the likely reliability of our coding scheme, and all articles underwent a quality control procedure involving a second coder to identify any mistakes or inconsistencies.

Double-Blind Extraction

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

Quality Control

All articles underwent the quality control procedure. Here, a second coder reviewed the data extracted from each article. 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 negative claim from the title/abstract has been extracted, 2) that any negative claim extracted was really a negative claim, 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

~~The primary analysis was descriptive.~~ The percentage of claims in each category across the titles, abstract population claims, abstract sample claims, and result texts are presented. 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. We used a Chi-squared test to test whether, if a "No Effect" interpretation was made in the results, it was more likely that a "No Effect" interpretation would also be made in the abstract than when a correct interpretation was made in the results. All data and code as well as our coding guidelines are openly available at https://osf.io/84puf/.

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.

Reliability and Quality Control

Double-Blind Coding

~~For 24 articles (8.5% of the total sample), two authors (BGF and the author originally assigned to the journal) extracted all the data independently of each other.~~ Only five of the article titles were identified as containing negative statements by either of the two coders, and from this, the two coders agreed on only one out of five (20%) of the articles about whether the title statement was about the tested sample or the population. Following discussion with the whole group, we agreed that it was often ambiguous whether the titles of articles were referring to the

specific sample tested or a wider population, and so we decided to combine these measures and have no sub-group analysis for the title claim, deviating from our original plan. When considering the category (Justified; Caveated, Similar or Ambiguous; No Effect) of the title claim, the two coders agreed on two out of six articles (33%). Three of the four disagreements occurred when one coder did not interpret the title as a negative claim, e.g. as in "Evidence that novel flavors unconditionally suppress weight gain in the absence of flavor-calorie associations" (Seitz et al., 2020), and one where a coder appeared to have made an error. From discussion within the group, it was evident that these ambiguous cases — where the statements were not clearly written as negative statistical results but involved an interpretation that did not directly corresponds to a specific non-significant result from the article — proved the most difficult during the whole coding procedure, and this affected the reliability of the title claims and population claims from the abstract.

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

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

Quality Control

Each article was checked by a quality controller. The initial coders identified 67 possible negative statements in the titles of papers, and the quality controller agreed with the classification of 39 (58%) of these statements, had a minor disagreement with six statements (9%), and a major disagreement with 22 statements (33%). Of note, 16 of these 22 major disagreements came from a single repeated error in which one individual coder coded 'ambiguous' for titles containing no negative statement. In the abstract, coders identified 281 negative statements about the specific sample tested in the paper. Of these, the quality controllers agreed with the classification of 250 (89%), had minor comments about 16 (6%), and major disagreements with 15 (5%). Coders identified a much smaller number of negative inferences about populations in the articles and disagreed more frequently: Of the 82 identified statements, the quality controllers agreed with the classification of 44 (53%), had minor comments about 18 (22%) and major disagreements with 20 (24%). Regarding the result texts from the article bodies, coders identified 282 results, of which the quality controller agreed with the classification and extracted *p*-value for 252 (89%), had minor comments for 13 (5%), and major disagreements for 17 (6%).

The quality control process allowed us to, i) identify any clear errors in the data extraction process, ii) highlight borderline cases where our coding scheme could not clearly categorize certain statements, and iii) assess the robustness of the coding procedure. In line with the results from the double-blind coding, the quality control process demonstrated a high inter-rater agreement and consistency with identifying and classifying negative sample statements from abstracts, and the corresponding results and *p*-values from the main text, yet greater inconsistency in deciding, i) whether titles and population statements were truly "negative" in the sense of being the result of a non-significant NHST, and ii) whether the authors were claiming the absence of an effect from these negative results. This inconsistency occurred mainly because many titles and population claims referred not to a certain statistical result but made a vague theoretical statement.

Title Claims

Forty-four titles contained negative 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 two groups or conditions being 'similar'. *Table 4* provides examples of these claims.

Abstract Claims

Abstract Sample Claims

We extracted 278 negative claims about a sample result. 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.

Abstract Population Claims

We extracted 63 negative claims about a population that followed on from the negative result within a sample. Of these, 45 (71%) were 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. Several of the classifications as ambiguous were due to authors' use of "main effect" when interpreting ANOVA terms, where we thought that saying there was "no main effect of X" was different enough to saying "no effect of X" to not be included in the "No Effect" category, although this highlights the somewhat arbitrariness of our categories. *Table 7* provides examples of the different types of result reporting.

Notably, if a sentence reporting the results in the results section was classified as "No Effect", it was more likely that a no effect interpretation would also be made 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 sample statements were classified as interpreting there being no effect. In contrast, of the 119 statements in the results classified as "non-significant", only 67 (56%) of the corresponding sample statements were classified as interpreting there being no effect. Nevertheless, "No Effect" interpretations in the abstracts were absolutely the most likely classifications for both "No Effect" and "Non-Significant" results statements.

p-value distributions

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

Error! Reference source not found. contrasts the distribution of *Figure 1* with the four simulated distributions of bodies of research performed where 80% of alternative hypotheses were correct, and studies had either 10, 33, 50 or 80% power to detect the true effect size of H1 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.

Discussion

We extracted and classified how animal cognition researchers reported the results of non-significant null hypothesis significance tests in 253 articles between 2019 and 2021. Across titles, abstracts, and results, we classified non-significant results as often being reported with the

"No Effect" phrasing that has often been labelled as erroneous (titles 84%; abstract sample results 63%; result text 41%). Reporting negative results as "Non-Significant" was less common in titles and abstracts, but as prevalent as "No Effect" phrasings in the results (titles 16%; abstract sample results 26%; result text 52%). The other, albeit less frequently classified method of reporting non-significant results was to comment on the similarity between two groups or conditions (abstract sample results 6%; result text 4%).

~~Overall,~~ these results demonstrate considerable heterogeneity in how animal cognition researchers report and potentially interpret non-significant results in published articles (according to our classification of these reports, made by other researchers in the field). However, it was often difficult to confidently categorize results due to the heterogeneity in how negative results were reported. are Nevertheless, our results suggest that negative 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. Possibly encouragingly, when researchers extended "No Effect" statements from the sample to the population, 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"[2]). Again, however, more research is needed to understand how such statements are interpreted and implemented by scientists and the wider community. 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.

Notably, the coding team found it difficult to identify and classify negative population statements in the abstracts of articles. This likely reflects the distance between the theoretical claims researchers wish to test and the actual statistical hypotheses that are tested, *i.e.,* rarely can

[2] Although we did not study this, it is likely this type of caveating is not unique to negative results but used to caveat positive findings, too.

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 "No Effect" interpretations 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 wrote "No Effect" interpretations 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. 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 researchers had interpreted as negative for the manual dataset, but *p*-values in the range .05-0.1 are often interpreted as "trends" or "marginally significant".

Conclusions

This study explored reporting and interpretation of negative result 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 non-significant results were often reported as if there were no differences observed in the sample, and this was the case in the titles, abstracts and result sections of papers, although it was most frequent in the titles and abstracts. Because of the distance between statistical hypotheses and theoretical claims, and uncertainty around how no difference statements are interpreted, the consequences of this putative error are uncertain. Nevertheless, these results suggest that researchers should pay close attention to the evidence used to support claims of absence of effects in the animal cognition

literature, and prospectively seek to, i) report non-significant results clearly, and ii) use more formal methods of assessing the evidence against theoretical predictions.

Acknowledgements

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

References

Aczel, B., Palfi, B., Szollosi, A., Kovacs, M., Szaszi, B., Szecsi, P., Zrubka, M., Gronau, Q. F., van den Bergh, D., & Wagenmakers, E.-J. (2018). Quantifying support for the null hypothesis in Psychology: An empirical investigation. *Advances in Methods and Practices in Psychological Science*, *1*(3), 357–366. https://doi.org/10.1177/2515245918773742

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

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

Beran, M. J., French, K., Smith, T. R., & Parrish, A. E. (2019). Limited evidence of number–space mapping in rhesus monkeys (*Macaca mulatta*) and capuchin monkeys (Sapajus apella). *Journal of Comparative Psychology*, *133*(3), 281–293. https://doi.org/10.1037/com0000177

Brecht, K. F., Müller, J., & Nieder, A. (2020). Carrion crows (C*orvus corone corone*) fail the mirror mark test yet again. *Journal of Comparative Psychology*, *134*(4), 372–378. https://doi.org/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. https://doi.org/10.1007/s10071-020-01425-9

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

Cunningham, P. J., & Shahan, T. A. (2020). Delays to food-predictive stimuli do not affect suboptimal choice in rats. *Journal of Experimental Psychology: Animal Learning and Cognition*, *46*(4), 385–397. https://doi.org/10.1037/xan0000245

DeVries, M. S., Winters, C. P., & Jawor, J. M. (2020). Similarities in expression of territorial aggression in breeding pairs of northern cardinals, *Cardinalis cardinalis*. *Journal of Ethology*, *38*(3), 377–382. https://doi.org/10.1007/s10164-020-00659-x

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

Fidler, F., Burgman, M. A., Cumming, G., Buttrose, R., & Thomason, N. (2006). Impact of Criticism of Null-Hypothesis Significance Testing on Statistical Reporting Practices in Conservation Biology. *Conservation Biology*, *20*(5), 1539–1544.

Fiedler, K., Kutzner, F., & Krueger, J. I. (2012). The long way from α-error control to validity proper: Problems with a short-sighted false-positive debate. *Perspectives on Psychological Science*, *7*(6), 661–669. https://doi.org/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. https://doi.org/10.1177/0959354312436870

Gelman, A., & Carlin, J. (2014). Beyond power calculations. *Perspectives on Psychological Science*, *9*(6), 641–651. https://doi.org/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. *Handbook on Quantitative Methods in the Social Sciences. Sage, Thousand Oaks, CA*, 389–406.

Goodman, S. (2008). A Dirty Dozen: Twelve p-value misconceptions. *Seminars in Hematology*, *45*(3), 135–140. https://doi.org/10.1053/j.seminhematol.2008.04.003

Guadarrama, S. S., Domínguez-Vega, H., Díaz-Albiter, H. M., Quijano, A., Bastiaans, E., Carrillo-Castilla, P., Manjarrez, J., Gómez-Ortíz, Y., & Fajardo, V. (2020). Hypoxia by altitude and welfare of captive beaded lizards (*Heloderma Horridum*) in Mexico: Hematological

approaches. *Journal of Applied Animal Welfare Science*, *23*(1), 74–82. https://doi.org/10.1080/10888705.2018.1562350

Harris, J. A., & Bouton, M. E. (2020). Pavlovian conditioning under partial reinforcement: The effects of nonreinforced trials versus cumulative conditioned stimulus duration. *Journal of Experimental Psychology: Animal Learning and Cognition*, *46*(3), 256–272. https://doi.org/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. https://doi.org/10.19227/jzar.v8i4.523

Hoekstra, R., Finch, S., Kiers, H. A. L., & Johnson, A. (2006). Probability as certainty: Dichotomous thinking and the misuse of p values. *Psychonomic Bulletin & Review*, *13*(6), 1033–1037. https://doi.org/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. https://doi.org/10.1163/1568539X-bja10018

Kawai, N., Nakagami, A., Yasue, M., Koda, H., & Ichinohe, N. (2019). Common marmosets (*Callithrix jacchus*) evaluate third-party social interactions of human actors but Japanese monkeys (*Macaca fuscata*) do not. *Journal of Comparative Psychology*, *133*(4), 488–495. https://doi.org/10.1037/com0000182

Koczura, M., Martin, B., Musci, M., Massimo, M. D., Bouchon, M., Turille, G., Kreuzer, M., Berard, J., & Coppa, M. (2021). Little difference in milk fatty acid and terpene composition among three contrasting dairy breeds when grazing a biodiverse mountain pasture. *Frontiers in Veterinary Science*, *7*, 612504. https://doi.org/10.3389/fvets.2020.612504

Kvarnemo, C., Andersson, S. E., Elisson, J., Moore, G. I., & Jones, A. G. (2021). Home range use in the West Australian seahorse Hippocampus subelongatus is influenced by sex and partner's home range but not by body size or paired status. *Journal of Ethology*, *39*(2), 235–248. https://doi.org/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. https://doi.org/10.1177/1948550617697177

Lakens, D. (2021). *Sample Size Justification* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/9d3yf

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., Baguley, T., Becker, R. B., Benning, S. D., Bradford, D. E., Buchanan, E. M., Caldwell, A. R., Van Calster, B., Carlsson, R., Chen, S.-C., Chung, B., Colling, L. J., Collins, G. S., Crook, Z., … Zwaan, R. A. (2018). Justify your alpha. *Nature Human Behaviour*, *2*(3), 168–171. https://doi.org/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. https://doi.org/10.1177/0959354311429854

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

Lilley, M. K., de Vere, A. J., & Yeater, D. B. (2020). Laterality of eye use by bottlenose (*Tursiops truncatus*) and rough-toothed (*Steno bredanensis*) dolphins while viewing predictable and unpredictable stimuli. *International Journal of Comparative Psychology*, *33*. https://doi.org/10.46867/ijcp.2020.33.03.01

Mayo, D. G. (2018). *Statistical inference as severe testing: How to get beyond the statistics wars*. Cambridge University Press.

Meza, P., Elias, D. O., & Rosenthal, M. F. (2021). The effect of substrate on prey capture does not match natural substrate use in a wolf spider. *Animal Behaviour*, *176*, 17–21. https://doi.org/10.1016/j.anbehav.2021.03.014

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. https://doi.org/10.1080/03610927608827392

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *231*(694–706), 289–337. https://doi.org/10.1098/rsta.1933.0009

Nord, C. L., Valton, V., Wood, J., & Roiser, J. P. (2017). Power-up: A reanalysis of "Power Failure" in neuroscience using mixture modeling. *Journal of Neuroscience*, *37*(34), 8051–8061. https://doi.org/10.1523/JNEUROSCI.3592-16.2017

O'Donoghue, E. M., Broschard, M. B., & Wasserman, E. A. (2020). Pigeons exhibit flexibility but not rule formation in dimensional learning, stimulus generalization, and task switching. *Journal of Experimental Psychology: Animal Learning and Cognition*, *46*(2), 107–123. https://doi.org/10.1037/xan0000234

Paijmans, K. C., Booth, D. J., & Wong, M. Y. L. (2021). Odd one in: Oddity within mixed-species shoals does not affect shoal preference by vagrant tropical damselfish in the presence or absence of a predator. *Ethology*, *127*(2), 125–134. https://doi.org/10.1111/eth.13110

Pereira, F. C., Teixeira, D. L., Boyle, L. A., Pinheiro Machado Filho, L. C., Williams, S. R. O., & Enriquez-Hidalgo, D. (2021). The equipment used in the SF6 technique to estimate methane emissions has no major effect on dairy cow behavior. *Frontiers in Veterinary Science*, *7*, 620810. https://doi.org/10.3389/fvets.2020.620810

Piefke, T. J., Bonnell, T. R., DeOliveira, G. M., Border, S. E., & Dijkstra, P. D. (2021). Social network stability is impacted by removing a dominant male in replicate dominance hierarchies of a cichlid fish. *Animal Behaviour*, *175*, 7–20. https://doi.org/10.1016/j.anbehav.2021.02.012

Pinto, P., & Hirata, S. (2020). Does size matter? Examining the possible mechanisms of multi-stallion groups in horse societies. *Behavioural Processes*, *181*, 104277. https://doi.org/10.1016/j.beproc.2020.104277

Ribes-Iñesta, E., Hernández, V., & Serrano, M. (2020). Temporal contingencies are dependent on space location: Distal and proximal concurrent water schedules. *Behavioural Processes*, *181*, 104256. https://doi.org/10.1016/j.beproc.2020.104256

Rose, E. M., Mathew, T., Coss, D. A., Lohr, B., & Omland, K. E. (2018). A new statistical method to test equivalence: An application in male and female eastern bluebird song. *Animal Behaviour*, *145*, 77–85. https://doi.org/10.1016/j.anbehav.2018.09.004

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225

Schino, G., Boggiani, L., Mortelliti, A., Pinzaglia, M., & Addessi, E. (2021). Testing the two sides of indirect reciprocity in tufted capuchin monkeys. *Behavioural Processes*, *182*, 104290. https://doi.org/10.1016/j.beproc.2020.104290

Seitz, B. M., Flaim, M. E., & Blaisdell, A. P. (2020). Evidence that novel flavors unconditionally suppress weight gain in the absence of flavor-calorie associations. *Learning & Behavior*, *48*(3), 351–363. https://doi.org/10.3758/s13420-020-00419-4

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

Vadillo, M. A., Konstantinidis, E., & Shanks, D. R. (2016). Underpowered samples, false negatives, and unconscious learning. *Psychonomic Bulletin & Review*, *23*(1), 87–102. https://doi.org/10.3758/s13423-015-0892-6

Vernouillet, A., Clary, D., & Kelly, D. M. (2021). Highly social pinyon jays, but not less social Clark's nutcrackers, modify their food-storing behaviour when observed by a heterospecific. *BioRxiv*, 2021.02.28.433225. https://doi.org/10.1101/2021.02.28.433225

Wu, Y., Petrosky, A. L., Hazzi, N. A., Woodward, R. L., & Sandoval, L. (2021). The role of learning, acoustic similarity and phylogenetic relatedness in the recognition of distress calls in birds. *Animal Behaviour*, *175*, 111–121. https://doi.org/10.1016/j.anbehav.2021.02.015

Yang, C., Tsedan, G., Fan, Q., Wang, S., Wang, Z., Chang, S., & Hou, F. (2021). Behavioral patterns of yaks (*Bos grunniens*) grazing on alpine shrub meadows of the Qinghai-Tibetan Plateau. *Applied Animal Behaviour Science*, *234*, 105182. https://doi.org/10.1016/j.applanim.2020.105182

Table 1(on next page)

Sources of articles containing negative results in their abstracts

1

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

2

Table 2(on next page)

Example categorization of sample-level statements

1

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.

2 .

3

Table 3(on next page)

Example categorization of population-level or title claims

1

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

2

Table 4(on next page)

Examples of claims in the titles of articles following non-significant NHST classified as “No Effect” and “Caveated or Similar”

1

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

2

3

Table 5(on next page)

Examples of claims about the sample in the abstracts of papers following non-significant NHST classified as “No Effect”, “Similar or Small Effect Size”, “Non-Significant” or “Ambiguous”

1

No Effect
N = 174, 63% "Levels of individuals sitting with their back to the window was unaffected by visitor number or noise." Hashmi & Sullivan, 2020 "The groups did not differ in their ability to follow human signals" Lazarowski et al., 2020
Similar or Small Effect Size
N = 17, 6% "Pair members demonstrated comparable responses towards a male 'intruder', as latency to respond and proximity scores were very similar between pair members in the majority of pairs examined" DeVries et al., 2020 "We found that individuals called back to sympatric and allopatric calls within similar amounts of time," Wu et al., 2021
Non-Significant
N = 71, 26% "Nutcrackers… did not significantly change their caching behaviour when observed by a pinyon jay." Vernouillet et al., 2021 "No significant correlations between degree of laterality and behavioral interest in the stimuli were found"

Lilley et al., 2020
Ambiguous
N = 16 (6%) “We also found no conclusive evidence that either the visual or the vibratory sensory modalities are critical for prey capture.” Meza et al., 2021 “No systematic variations on space allocation were observed in neither experiment” Ribes-Iñesta et al., 2020

2

Table 6(on next page)

Examples of claims about populations in the abstracts of papers following non-significant NHST classified as “No Effect” and “Caveated, Ambiguous or Similar”

1

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

2

Table 7 (on next page)

Examples of statement reporting the results in the results sections of papers using non-significant NHST classified as “No Effect”, “Similar or Small Effect Size”, “Non-Significant” or “Ambiguous”

1

No Effect
N = 113 (41%) During farrowing, No Effect of the treatments was seen on the percentage of time spent (3.22 % vs. 1.90 %, P = 0.372) on the nest-building behaviour" Aparecida Martins et al., 2021 "There were no differences between treatments in the frequency or duration of birds flying between walls" Stevens et al., 2021
Similar or Small Effect Size
N = 12 (4%) "The average time yaks spent grazing was similar among shrub coverage groups (P = 0.663)" Yang et al., 2021 "The number of sessions required to reach criterion didn't reliably differ between groups" O'Donoghue et al., 2020
Non-Significant
N = 140 (52%) "Comparing the pooled data of all crows, no significant increase in the number of mark-directed behaviors during the mirror mark condition was found compared with the no-mirror sham condition." Brecht et al., 2020 "There was no significant effect of removal type on changes in display strength in either dominant males or subordinate males."

Piefke et al., 2021
Ambiguous
N = 10 (4%) "As can be seen in Figure 1D, there was no difference in response rates after R 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

2

3

Figure 1

Distribution of non-significant p-values from result sections of 198 articles in animal cognition and related fields, with a density distribution overlaid in pink.

The dotted line shows the average density.

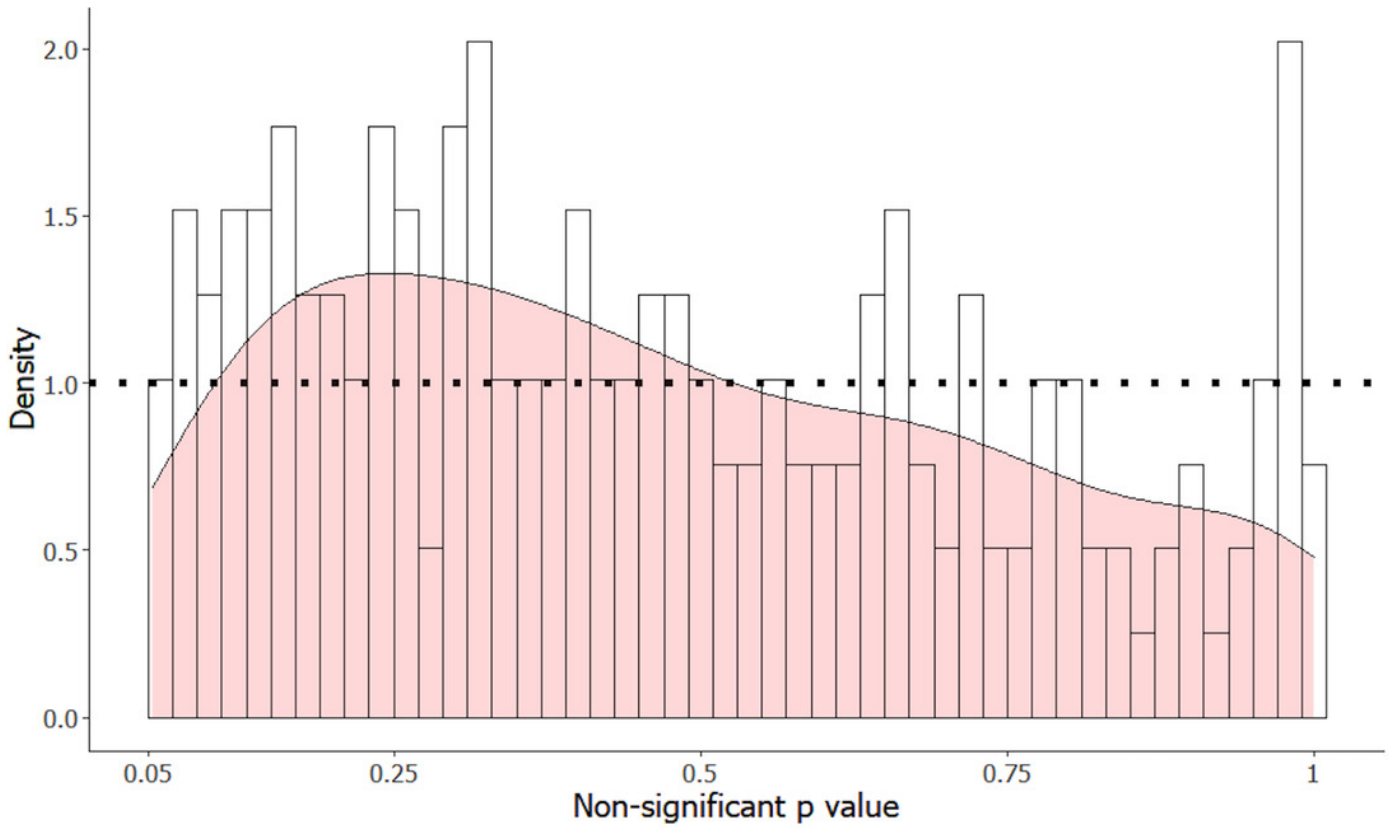

Figure 2

The observed p-value distribution of 198 p-values > .05, compared to 3 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 3 simulated distributions where 80% of alternative hypotheses were correct, with studies performed at either 10%, 33%, 50% or 80% statistical power.