# **Null Hypothesis Significance Testing in Animal Behaviour Research**

Contents

[The Project 2](https://d.docs.live.net/41e7a4d6e454be82/Documents/PhD/aaThesis/5_NHST/Stage1Guidelines.docx#_Toc48748705)

[Aims 2](https://d.docs.live.net/41e7a4d6e454be82/Documents/PhD/aaThesis/5_NHST/Stage1Guidelines.docx#_Toc48748706)

[Procedure. 2](https://d.docs.live.net/41e7a4d6e454be82/Documents/PhD/aaThesis/5_NHST/Stage1Guidelines.docx#_Toc48748707)

[Coding Guidelines 3](https://d.docs.live.net/41e7a4d6e454be82/Documents/PhD/aaThesis/5_NHST/Stage1Guidelines.docx#_Toc48748708)

# 

# **Project Overview**

## Background

· False-positive results have been a heavy focus in the scientific reform literature.

· False negative results are also concerning, particularly in areas where low power research might be common – such as animal behaviour research.

· Null hypothesis significance testing and *p*-values are easily misunderstood and mis-used, and it is plausible that researchers might often accept null hypotheses without this being warranted.

· e.g. Aczel et al. (2018, p. 357) “*examined all nonsignificant findings mentioned in the abstracts of the 2015 volumes of Psychonomic Bulletin & Review, Journal of Experimental Psychology: General, and Psychological Science (N = 137). In 72% of these cases, nonsignificant results were misinterpreted, in that the authors inferred that the effect was absent. Second, a Bayes factor reanalysis revealed that fewer than 5% of the nonsignificant findings provided strong evidence (i.e., BF01 > 10) in favor of the null hypothesis over the alternative hypothesis.”*

## Aims

· Estimate the prevalence of accepting the null in animal sciences (behaviour, cognition, learning and welfare)

· Produce a resource that can **improve** statistical inference and report in these fields.

## 

## Procedure

1. **Review coding guidelines** document and flag any uncertainties or disagreements. *COMPLETED*

2. **Pilot study**: 6 articles to be coded. *COMPLETED*

3. **Group feedback**, protocol finalized. *COMPLETED*

4. **Data Collection**: Coders to each be assigned one or two journals (until all have been assigned), and to code through this journal (back from the most recent article) until 15 articles using NHST and reporting negative results in the abstract/title have been coded.

5. **Quality Control**: A second coder will verify the first coders decisions. This coder will not be blinded, and will check that, i) the negative claim is actually a negative claim, ii) the coded NHST does match with the abstract/title claim, and iii) whether they agree with the categorisation of the claim. Disagreements will be flagged and judged by a third coder in discussion with the original coder and quality controller.

*Journal pool (20 journals):*

*Animals, Animal Cognition, Animal Behaviour, Animal Behavior and Cognition, Applied Animal Behaviour Science, Behaviour, Behavioural Processes, BioRxiv: Animal Behaviour and Cognition, Ethology, International Journal of Comparative Psychology, Journal of Applied Animal Welfare Science, Journal of Comparative Psychology, Journal of Ethology, Frontiers in Psychology: Comparative Psychology, Frontiers in Veterinary Science: Animal Behaviour and Welfare, Journal of Experimental Psychology: Animal Learning and Cognition, Journal of Zoo and Aquarium Research, Learning and Behavior, PeerJ: Animal Behaviour, Peer Communities In: Animal Science*

# 

# **Coding Guidelines**

For each paper:

1. Screen the **abstract** for any negative statements about, i) **the results in the current sample**, and ii) **the population.**

Sample results examples:

“dogs did not perform differently to wolves”

“there was no effect of X on the dogs’ performance”

“dogs’ behavior was not significantly different between time A and time B”

Population inferences examples:

“this suggests that dogs are not different from wolves”

“showing that X does not affect dogs”

“suggesting that dog behavior does not change over time”

If there are no negative statements, move onto the next paper. Else…

1. Record the paper’s information (title, first author, journal and year) and the negative statement that you think is most clearly related to the paper’s main claim. If present, record both the sample and population statement on separate lines, but only record one of each per paper.
2. Check whether the paper reports the results of null hypothesis significance tests. If you’re uncertain, code yes if the paper presents *p*-values. If no, move onto the next paper, else if yes…
3. Find and record the corresponding statistical test in the results (e.g. “dogs did not perform significantly differently in condition A (25±5) than in condition B (28±5, *p* > .05)). If there are multiple p-values within a table that apply to the statistical result, record the one that you think is most central to the claim - if they are all equally important to the claim, report the first one presented.

If there is more than one corresponding statistical test within an experiment, record the test result that you think is most relevant to the claim. If the claim is supported by multiple experiments, record the test result from the first experiment presented.

1. Categorize the interpretation of the **sample** results given in the **abstract and results text** according to the following categories:

a) Correct: Reports that there was no *significant* difference between two conditions, or words to that effect (see Table 1 for examples), OR any statement that is literally correct, e.g. Dogs did not perform higher in condition A (mean = 5) than condition B (mean 7).

b) No 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.

c) Ambiguous: A statement that neither suggests that the results were the same, nor that there was no significant difference.

Table 1 provides example statements for each category.

**Table 1**

Example categorization of sample level claims

|  |  |  |
| --- | --- | --- |
| **Correct**    Reports that there was no *significant* difference between two conditions, or words to that effect. | **No 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. | **Ambiguous**    A statement about the results that neither suggests they were the same, nor that there was no significant difference. |
| There was no significant/detectable difference between X and Y.  We did not detect a difference between X and Y (or any other statement implying failing to find a signal within noise).  We did not find a significant effect.  X was not significantly related to Y.  X did not perform significantly above chance.  X performed significantly above chance, but Y did not.  There were no significant differences between X and Y’s performance.  X did not do A more in condition Y than condition Z (and this is genuinely true in the data – see Note). | There was no difference between X and Y.  There was no effect.  There was no evidence of an effect.  There was no relationship between X and Y.  We did not find/observe/see a difference between X and Y.  We did not find an effect.  We found no evidence of an effect.  X performed at chance levels.  X performed above chance, but Y did not (if Y > chance, but not significantly).  X and Y performed equally.  X did not do A more in condition Y than condition Z (but this is not true in the data).  We did not find/observe/see a relationship between X and Y.  We did not observe X performing above chance.  We found that X performed above chance, but Y did not.  We did not find/observe/see a difference between X and Y’s performance. | X and Y were similar.    There was no large/clear difference between X and Y. |

**Note.**If something is literally true, then it should always be coded as correct. Example 1, if the statement was “Dogs looked at the target for the same amount of time in Condition A (10.0 seconds) as in Condition B (10.0 seconds, *p* = 1)” this should be coded as correct - even though they wouldn’t have spent the *exact* same time, we will code this relative to the precision they report, e.g. 1 decimal place in Example 1. Example 2, “Dogs did not look longer at Target A (mean = 7 seconds) than Target B (mean = 10 seconds, p = 0.8”. Here, the dogs looked for less time at Target A than Target B, and even though it was not significantly less, this means that the statement “dogs did not look longer at Target A than Target B” is true.

7. Categorize the interpretation of the **population** results given in the abstract according to the following categories:

a) Correct: Comments on statistical power, uses equivalence tests or otherwise justifies why 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

b) Caveated: Interprets the non-significant results as suggesting/indicating etc. that X and Y do not differ in the population

c) No Effect: Interprets the non-significant result as showing that X and Y do not differ in the population

Table 2 provides example statements for each category.

**Table 2**

Example categorization of population level claims

|  |  |  |
| --- | --- | --- |
| **Correct**  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. | **Caveated**  Interprets the non-significant results as suggesting/indicating etc. that X and Y do not differ in the population. | **No Effect**  Interprets the non-significant result as showing that X and Y do not differ in the population. |
| 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.      Any statement about the test likely being low powered and this making it difficult to interpret what the results mean at the population level. | …Suggesting that X 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. | … 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. |

8. If the title contains a negative statement about **the sample**, categorize this according to Table 1 too,

9. If the title contains a negative statement about **the population**, categorize this according to Table 2 too,