Perspective

Beyond the null: Recognizing and reporting true negative findings

Manon K. Schweinfurth[1,*] and Joachim G. Frommen[2]
[1]School of Psychology and Neuroscience, University of St Andrews, St Andrews KY16 9JP, UK
[2]Department of Natural Sciences, Manchester Metropolitan University, Manchester M15GD, UK
*Correspondence: ms397@st-andrews.ac.uk
https://doi.org/10.1016/j.isci.2024.111676

SUMMARY

Science is based on ideas that might be true or false in describing reality. In order to discern between these two, scientists conduct studies that can reveal evidence for an idea, i.e., positive findings, or not, i.e., negative or null findings. The outcome of these studies can either be *true*, i.e., reflecting the real world, or *false*. Much has been said about disentangling true from false positive findings and the danger of a publication bias toward positive findings. Here, we argue that publishing negative findings is important to provide an accurate picture of the real world. At the same time, we highlight that a cautious approach should be taken to minimize the impact of publishing *false* negative findings, which has received limited attention so far. We discuss sources of false negative findings, using experimental and observational animal behavior and cognition studies as examples, which often differ from those of false positive findings. We conclude by recommending strategies for rigorous studies, such as conducting positive controls, selecting diverse samples, designing engaging protocols, and clearly labeling negative findings. These practices will lead to studies that contribute to our knowledge, regardless of whether they result in positive or negative findings.

"Absence of evidence is not evidence of absence"

Carl Sagan[1]

Imagine a scenario that is too well known to many scientists. You had a fantastic idea, you designed the study, you got funding, you received ethical approval, you conducted the study, and the predicted effect is not there. Without doubts, this is a frustrating experience.

In the past, many such negative findings were not published, leading to a publication bias in the current literate, where publications report more positive findings than expected across disciplines.[2,3] To tackle this underrepresentation in the literature, new journals were established that focused on publishing negative findings, e.g., the *Journal of Negative Results* or the *Journal of Negative Results in Biomedicine*. Today, some leading journals explicitly encourage publishing negative findings (e.g., *Nature Human Behavior* and *PLoS ONE* with examples such as[4,5]). As a result, the *Journal of Negative Results in Biomedicine* ceased publishing in September 2017 because they felt their mission was successfully completed.[6]

We highly welcome this development. Both positive and negative findings need to be reported to provide an accurate representation of the research field, rather than focusing merely on sensational or newsworthy findings. However, in order to be helpful in advancing the scientific field, negative findings need to be founded on a reliable body of evidence. Indeed, just like positive findings, negative findings might be true, when there is

indeed no effect, or false, when there is an effect that was not detected (Figure 1A). Much has been said about the danger of publishing false positive findings[7] and mitigation techniques have been implemented over the past years. Current open science practices, such as preregistering protocols and analyses prior to data collection as well as presenting peer review and raw data files openly, are aimed at detecting and reducing false positives in the literature. Yet, the pitfalls of publishing trustworthy negative findings have received less attention although they cannot be easily solved by following open science techniques. This oversight might be based on the perception that negative findings are more likely to be true because confirmatory biases or p-hacking can be excluded. Indeed, negative findings have become more prevalent in recent years, yet two out of three psychology articles reporting non-significant results contain evidence of at least one false negative[8] and in >70% negative findings were misinterpreted.[9] False negative findings can have dire consequences. For example, >30% of such erroneous claims in Educational Psychology could be linked to misguided educational theory, practice or policy.[10]

Therefore, the aim of this article is to explore challenges and solutions for negative findings. We first highlight reasons for why negative findings might not be conclusive evidence for the absence of an effect, which might be different from distinguishing between true and false positive findings. We use examples from behavioral and cognitive research, which we will discuss for observational and experimental studies separately. While we focus on the fields where our own expertise is strongest, these issues can be generalized to other disciplines. Second,

A | True and false findings

Real effect is...

		present	absent
Measurement shows effect	**yes**	**true positive** also known as: 1-β, true hit, true detection	**false positive** also known as: Type I Error, α, false alarm/hit, over-detection
	no	**false negative** also known as: Type II Error, β, false miss/rejection, under-detection	**true negative** also known as: 1-α, true rejection, true exclusion

B | Sources of false findings and how to prevent them

Sources	Solutions
• Small sample size	• Test large and diverse samples
• Unrepresentative subjects	• Use diverse situations
• Task not valid	• Discuss & pilot design
• Lack of motivation	• Put yourself into your subjects' shoes
• Complex task & poor task understanding	• Add positive control
• Experimenter effects	• Habituation
• Low detection rate	• Test for equivalence
• Level of analyses inappropriate	• Consider all test outcomes before study start

Figure 1. Navigating false negatives
Studies result in true or false positive or negative findings, whereas synonymous terms are used in different fields (panel A). False negative findings can result from various sources, which can be prevented or mitigated (panel B).

we will expand on why false negative findings are problematic and when negative findings should be reported. We conclude by making recommendations that will result in more conclusive findings, be they positive or negative.

REASONS FOR FALSE NEGATIVE FINDINGS IN EXPERIMENTAL STUDIES

Negative findings are not necessarily conclusive evidence for the absence of an effect because there can be various reasons why this seemingly negative (or null) finding might be false. Here, we provide a summary of the most common sources that we have encountered during journal clubs, peer reviews, and editorial work.

(1) **Sampled individuals might be too few.** The chance of detecting an effect depends on the effect and sample size. Studies that rely on only a few individuals will have low power to find an effect of low to medium size.[11–13] Especially measurements that are characterized by high variation require large sample sizes to detect an effect. For example, a systematic literature review revealed that only 10–20% of studies conducted in the field of Animal Behavior exceeded a statistical power of 80%, i.e., the probability of obtaining a significant result when the effect is true.[14] Additionally, studies are often biased toward a small pool of species,[15] limiting generalizability of effects across phylogeny. It is hence unclear if a study reports no effect because there is no effect, which would be a true negative finding, or whether it lacks the power to detect an effect, which would be a false negative finding.[13,16]

(2) **Sampled individuals might not be representative.** Some research fields are biased toward certain individuals especially when working with long-lived species, as

these individuals are often probed repeatedly in different studies, leading to non-independent samples.[15] It is important that test subjects – especially when working with few individuals of a given species repeatedly - are representative for their species or at least population, which is not trivial to assess.[17,18] For instance a multi-site comparison of primate working memory abilities showed that individuals of the same species that were living at different sites performed differently, which might be the result of different levels of experiences with such tasks.[19] In addition, it was found, for instance, that a lack of predation and parasites in the lab boosts the immune response of guppies (*Poecilia reticulata*), which changes their behavior.[20] Hence negative findings on a sample might not be generalizable for the entire species.

(3) **The study protocol may not adequately assess the concept it claims to test.** Scientific studies that make use of non-human subjects are necessarily planned from a human perspective. However, study animals might perceive the world in a fundamentally different way. Hence, while the experimenter designs a protocol that can be validated by human subjects, the set-up might not be meaningful for the test animals. For example, many fish species see light in the ultraviolet spectrum[21] and use ultraviolet coloration to communicate during mate choice[22,23] and social interactions.[24,25] As human researchers lack this ability, we tend to design studies excluding UV light. Finding no evidence for the use of visual cues in certain interactions of fishes might therefore be because not all necessary visual cues were available to them.

In addition to perceptual differences, species differ in their likelihood to perform behaviors under some social settings. For example, many studies have aimed to test whether one of our closest living relative, the chimpanzee (*Pan troglodytes*), shows evidence for Theory of Mind, which is the ability to ascribe mental states to others like "I know what you know."[26] After years of research and using different protocols, it had been concluded that chimpanzees do not have a Theory of Mind.[27] This false negative finding was revealed by an innovative study, which used a competitive rather than a cooperative setting, demonstrating that chimpanzees after all are able to know what others know.[28] Probably, the discovery was hindered as humans frequently apply the Theory of Mind in cooperative settings.[29] Hence, older studies did not test the *general occurrence* of Theory of Mind, as intended by the researchers, but instead Theory of Mind *under very specific conditions*. Nowadays, it is well established that chimpanzee have a Theory of Mind if tested in the right setting.[30].

(4) **The study design might not be motivating enough for individuals to respond.** Many studies provide a form of stimulus to elicit an expected result in the test subjects. Such stimuli might for example include the presence of conspecifics[31] or food.[32] Yet, designing a task that is motivating for a range of test subjects with different backgrounds is not trivial. For example, while food is used

regularly as a reward for successful trials, the value of food types differs for individuals based on hunger levels,[32,33] individual preferences[34] cultural differences,[35] or previous experience.[36,37] Hence, not all test subjects might be equally motivated to partake, leading to noise in the data. If the motivation is too low, subjects might not respond at all or only some individuals will respond. Additionally, if the motivation to partake is too high, attention to details and impulsivity control might be reduced,[36] also leading to false negative findings.

(5) **The study design might be too complex to understand.** For certain questions, for instance to rule out previous experience, it can be important to expose subjects to novel (and hence often artificial) tasks, settings, or objects. However, this can be unintuitive for participants, especially when participants are not human.[38] For instance, chimpanzees are more likely to be prosocial toward their conspecifics in less complex and more naturalistic experimental tasks.[39] Similarly, the performance in a spatial navigation task of three-spined stickleback (*Gasterosteus aculeatus*) and minnow (*Phoxinus phoxinus*) has been shown to be dependent on task complexity,[40] exemplifying the risk that overly complicated mazes lead to the erroneous assumption of the absence of certain cognitive abilities in these species. Hence, to conclude that an effect is absent requires experimental proof that the test animal understood the task. Yet, not all studies reporting negative findings demonstrate successful task-understanding controls (reviewed in[41,42]).

(6) **Experimenters or environments might prevent individuals from responding according to their natural or best performance by disturbing or distracting them.** Lab studies often include interactions with the test subjects, which can include catching, carrying, or placing them into a testing arena. All of these interactions can induce stress and affect their behavior,[43] bearing the risk of confusing a true negative finding with a test outcome caused by neophobia or fear. To avoid anxiety, researchers use habituation techniques, in which test subjects are exposed to test conditions prior to the test. For example, spotted rainbowfish (*Melanotaenia duboulayi*) that were habituated to the study protocol and testing arena show better escape responses toward novel trawl apparatuses compared to when they were not habituated.[44] However, habituation is not a straightforward solution and contains a strong species- and setup-dependent component.[45] For example, agonistic behaviors of social cichlid fish shown toward a mirror decrease when they got habituated to the experimental setting,[46] probably because mirror images cannot harm test subjects in contrast to real opponents.[47] Consequently, several species, including fishes, eventually learn to use their mirror image to inspect their own body instead of fighting it.[48,49] In such cases exposing the test individual to the experimental setting for too long might be the cause for false negative findings if one is interested in aggression. Yet, some skills can only be revealed after extensive training and habituating phases, such as chimpanzees

showing evidence to understand human words after years of language training[50]

Sub-optimal acclimation is not the only source of creating false negative results. Other habituation effects impact study results, too. Carrion crows (*Corvus corone*) and ravens (*Corvus corax*), for example, are more likely to participate and succeed in a cognitive task with a familiar compared to an unfamiliar experimenter,[51] highlighting the role of a relationship between experimenters and test subjects. In contrast, switching experimenters during a study on Montagu's harrier (*Circus pygargus*) chicks decreased stress levels and aggressive behaviors.[52] In addition to experimenters, social environments can impact subjects. For example, highly social orange-winged amazons (*Amazona amazonica*) are more likely to participate in a behavioral task when tested in a social compared to an individual setting.[53] All these inter- and intraspecies social effects can make it more difficult for test subjects to perform well and hence might lead to false negative findings.

REASONS FOR FALSE NEGATIVE FINDINGS IN OBSERVATIONAL STUDIES

Observational studies differ from experimental studies in that they usually interfere less with the study subjects and often happen under less controlled conditions. As a result, false negative findings can be caused by factors that differ from lab studies, which we will discuss below. Publication bias toward positive findings is difficult to assess but might be more pronounced in observational compared to experimental studies because observational studies are often based on multiple observations instead of a few pre-defined responses to a certain task in experiments. As a result, it is unlikely that one picks a negative over a positive finding as the basis for a publication. Furthermore, findings from observational studies can rarely be fully replicated due to social and ecological changes in the test population or population differences[54] and hence the frequency of false positive and negative findings are hard to estimate. Nonetheless, the absence of a trait can be an important finding, and we encourage reporting them, as long as it is unlikely that the finding might be false.

(1) **The sample size might be insufficient.** Although observational studies are often based on larger sample sizes than experimental studies, some patterns can only be revealed when studying a large number of individuals of a species. For example, only when the behavior of chimpanzee groups all over Africa was compared, traits were revealed that suggested that non-human animals have culture.[55] Hence, comparable to under-powered lab studies, conclusions about the absence of traits from non-representative samples should be drawn with caution.

(2) **The sample might not be representative.** While it is often feasible to directly observe all group members in captive settings, this is more difficult in the wild. Personality and rank differences can affect habituation and

hence observability of certain group members.[56] Furthermore, it is important to ensure that not only conspicuous groups or traits are studied, as this can bias the literature.[57] Likewise, populations differ, and what is found in one population might not be generalizable to the entire species. Carrion crows, for example, are described as pair-breeding throughout most of their range. Still, there are populations in which breeders regularly accept brood-care helpers at their nest, challenging pair-living as the only social structure in this species.[58] Unrepresentative samples might therefore lead to false negative findings that cannot be generalized to the entire species.

(3) **The detection rate to demonstrate an effect might be too low.** In order to exclude false negative findings, subjects need to be studied under diverse circumstances, including different temporal (like the entire time of the day and season) and spatial resolutions (like the entire ecological niche of the individual and species). For example, honey-dipping tools were only discovered in West African chimpanzees after using direct (camera traps) and indirect (collection of abandoned tools) observational techniques over 23 consecutive months in four different communities.[59] Further, the observation length can impact detection rates. For example, decades of data collection over multiple generations can be necessary to relate life-history data to behavioral traits and cognitive skills.[60,61] Similarly, to analyze patterns in rare behaviors, long-term datasets are needed. For example, 40 years of research were needed to document rare cases of chimpanzee mothers carrying their long deceased infant, allowing for suggestions about the mechanisms and functions.[62] Finally, observer presence can disturb the natural behavior of individuals and hence reduce detection rates of certain skills or behaviors, as shown in Colombian white-faced capuchin (*Cebus capucinus*).[63]

(4) **False negative findings might be a result of masking effects and missing variables.** Studying animals in their natural environment has the benefit of measuring behaviors that are meaningful to the subjects in their natural context. However, a drawback is that it is difficult to focus on single effects, as there is a virtually endless number of potential co-effects and confounding factors. Some effects can be easier to measure than others and hence may mask relationships with the latter that are harder to detect. For example, a simulation study on cooperation networks demonstrated that provided help can be explained by kinship and reciprocity, but because nepotism can be more easily and reliably detected, it often masks the effect of reciprocity on helping decisions, especially when the sample size is small.[64] Hence studies with a small sample size might conclude that an effect is of minor importance or absent while it is merely masked by a co-effect and hence a false negative finding.

In addition to masking effects, missing variables can lead to false negative findings. Relationships are almost always a result of several factors and hence they might

not be easily detectable without knowledge of more than a few parameters. For example, in several territorial vertebrate species territoriality can only be meaningfully observed if several parameters are measured at the same time, e.g., intruder displacement in addition to site fidelity.[65]

REPORTING NEGATIVE FINDINGS

There is little doubt that the scientific endeavor strongly benefits from the publication of true negative findings. Not publishing true negative findings can impact meta-analyses by overestimating effect sizes,[66] and can lead to the establishment of false positive facts.[67] This can be illustrated by the following example: If a study has been conducted 20 times and only one repeat resulted in a positive finding, it is most likely a statistical artifact. However, this artifact could not be detected as such without the publication of at least some of the other 19 studies, which put the one positive finding into context.[68]

Given this risk, one could argue that it is always important to publish negative findings, regardless of whether true or false, because they may contain some information. This implies that there is always something to learn from negative findings, even if it is just an instruction of "how to not do it."[69] The problem is that there are often multiple reasons why something did not work out. As a result, it is almost impossible to assess which part of a study "not to do" in the future, especially if the study was not carefully executed in the first place. For example, a study that did not result in any behavioral change can have multiple reasons. The setup might not be valid, or it is valid, but the test subjects were unrepresentative, the training was inappropriate, task understanding was poor, and so forth. On top and more difficult to detect, many studies need considerable knowledge about the model species, handling techniques, or the experimental set-up which researchers only gain with experience. In such scenarios, learning how to conduct a study or observe certain traits might require extensive training and can affect study results.[70,71]

We think a far better approach than publishing every finding is to empower researchers to distinguish between true and false negative findings and publish only true findings in academic journals for several reasons (Figure 1B). First, false negative findings can have detrimental effects - sometimes even more than false positive findings. For example, a meta-analysis suggested that studies in Animal Welfare that aimed to assess the adversity of events, such as transporting live animals, are often underpowered and hence have a higher chance of resulting in false negative findings with potential severe wellbeing implications.[72] Second, if every study, whether true or false, is published, it will be difficult to discriminate between these findings, known as the "cluttered office" effect.[73] In 2016, it was estimated that two articles per minute were published - in the biomedical sector alone.[74] The more studies are available, the more difficult it is to keep track of every publication and the less time each researcher has available to assess their validity, which is especially problematic when the article covers a species or discipline that the reader is less familiar with. Third, publishing every dataset means that more false (positive and

negative) findings are published, which impacts the overall quality of science.[75] This does not necessarily mean that findings that bear the risk of being false could not be published at all. They might be submitted to specialized journals[69] or online platforms to avoid any confusion with true findings. However, it should be emphasized that no firm scientific conclusions should be drawn from such studies.

Therefore, we advocate that only those negative findings should be reported in peer-reviewed journals that provide evidence that they are likely to be true, for which we give concrete recommendations below. The file-drawer effect, which describes a systematic bias in scientific research to not publish negative findings, is harmful to science. However, there are good reasons to file studies that bear the risk of being false, and it needs expert knowledge to discriminate between such studies rather than indiscriminately publish positive and file negative findings. Better than filing any studies, however, is to design rigorous studies that control for common reasons why negative findings might be false. This is not only beneficial in terms of costs and time, but also avoids exposing subjects unnecessarily to experimental manipulations or repeated observations.

RECOMMENDATIONS

When designing a study, scientists tend to think about alternative explanations for positive findings and conduct control experiments to rule these out. While this is good practice, we encourage scientists to also think about alternative explanations for a negative finding and how one could be confident that a negative outcome reflects true effects. Many best-practice recommendations emphasize the danger of false positive findings with a strong focus on minimizing them. Open science practices are most effective in reducing the likelihood of false positives by preventing p-hacking or data dredging, for instance. However, while reducing the number of false positive findings is crucial, false negative findings are also an issue that must not be overlooked. While open science practices primarily focus on mitigating false positive findings, openly deposited methods, analyses, and data can also reveal positive effects that were missed, thereby contributing to identifying false negatives. Nevertheless, open science practices alone cannot fully address the risk of false negative findings. In the following, we list some suggestions that are relevant before, during, and after conducting a study and go beyond open science practices. Note that this list is not exhaustive and is aimed at stimulating further thoughts dependent on the research question, study species, and testing environment.

Conduct positive controls to ensure the study tests what it claims. If the study revealed a negative result, it is important to demonstrate that the treatment itself worked. It might be that the treatment is insignificant for test subjects, they are distracted, or not motivated to respond. For this, it is important to include a positive control, which ensures the experimental setup is functioning as intended. Such a control can rule out many possible confounding effects and may take various forms. It can replicate earlier findings or show an expected response by the test subjects toward the stimuli. Specifically, it shows that the

experimenter and the setting can produce meaningful data and that subjects perceive the stimuli and are motivated to respond. For example, a study on magpies (*Pica pica*) investigated whether the birds are attracted to shiny objects.[76] Birds were tested in captivity and in the wild, but none of the birds was attracted to shiny objects. A positive control condition revealed, however, that the birds paid attention to provided items and readily picked up food that was next to the objects. Hence, the researchers could demonstrate that the set-up itself worked, but that contrary to common beliefs magpies appeared to show no significant attraction to shiny objects. Furthermore, and especially in rather artificial tasks, it is crucial to test for full task understanding by setting up probing trials and including only those subjects in the test that passed them. For example, a study with an artificial food-provisioning task revealed reciprocal cooperation in chimpanzees only in those who passed the task understanding controls.[77] Finally, we think it is important to share and discuss the study protocol with peers, which can be done in seminars, at conferences, or via pre-registrations. Receiving feedback and discussing alternative explanations is much more productive before than after conducting a study, which is the case for peer-review because the protocol can be easily refined at an earlier stage. Here, discussions should range from all possible study outcomes, positive or negative, to seemingly minor protocol details, like randomizations and participant selection, to cover ideally all aspects of the study. The ARRIVE guidelines provide a good starting point and their application has been suggested to improve the quality of studies.[78]

Aim for appropriate sample sizes in order to provide stronger evidence for a negative finding. It is part of science that ideas turn out wrong and no effect of a treatment can be found. In this case, it is important to assess whether the lack of a treatment effect is based on a study that lacks the power to detect a true effect - false negative finding - or whether test subjects show indeed a similar response to the treatments - true negative finding. Depending on whether one follows a Frequentist or Bayesian approach, the options differ.[79] Following the Frequentist approach, equivalence tests should be conducted to assess whether responses are indeed equal,[14] confidence intervals should be used to provide parameter estimates with degrees of certainty,[80] and power should be calculated to assess whether the sample size is too small to detect a difference.[81–83] Alternatively, the Bayesian approach enables a more nuanced understanding of the data, including a level of confidence, using Bayes Factors,[84] and more direct support of no differences between two treatments,[85,86] especially when dealing with small sample sizes.[87]

Use representative samples and report their background. During data collection, it is important to generate meaningful and valid data for which a diverse pool of subjects is needed that show their undisturbed and representative behavior. Individuals differ and the more a researcher reduces or "standardises" the subject pool and their environment, the less generalizable are the study outcomes, which can lead to poor reproducibility.[17,88,89] It is therefore crucial to keep the source of variability in mind (see[18] for detailed discussion). For example, it can make a difference whether individuals are related (or even inbred), had early life experiences that make them aversive to a task, have

already been tested and therefore have been influenced by other experiments, or are bonded to the human observer (see above). Also, different rearing and keeping conditions have to be kept in mind when concluding that a result is a true negative.[90,91] While some variation can be achieved by simply keeping animal subjects under more diverse conditions,[92] other factors, such as social background and rearing history, are admittedly more difficult to diversify.[18] This can be achieved by collaborations between labs.[93,94] Likewise, before concluding that a certain trait or behavior is absent, it is crucial to ensure that this absence is not just based on population differences or can be explained by limited chances to observe a behavior in a population.

Make sure the study protocol is engaging for individuals. Before conducting any study, it is important to design a set-up that is motivating and relevant to the test subjects. To do so, it might be helpful to put oneself into the shoes of a subject and consider what they perceive and encounter in their given environment, how they would react naturally, and whether the stimuli would matter to them in their normal life.[95] For example, despite many years of research, there was no convincing evidence that non-human primates would understand false beliefs. Only after exposing them to a human in an ape costume, this skill could be revealed.[96] In addition, by testing animals in their natural environment or by mimicking real-world circumstances, ecological validity and hence relevance can be increased.[97]

Consider whether the individuals had a fair chance to show the expected results. Using complimentary testing protocols can provide insights into whether subjects are unable to solve a task generally or just in a certain setting.[98] In addition, caution should be paid to habituating individuals to testing environments or the presence of an observer. Just like learning criteria, experimenters should *a priori* define thresholds that suggest appropriate habituation while avoiding overstimulation to gather valid data (see above). Further, one has to ensure that detection rates are high to observe the trait of interest. This can be achieved by testing individuals under different spatial and temporal resolutions,[99] using extended periods of observer presence[100] or using novel technology to avoid the presence of observers.[101] Technical advancements have created exciting and less time-consuming opportunities to observe individuals over time and spatial scales, which have so far been impossible to explore via direct observations. This has resulted in astonishing studies from tiny hoverflies migrating at high altitudes[102] to giant Humboldt squids (*Dosidicus gigas*) living in the deep sea.[103] Still, it is important to note that not all data can be remotely collected for extended periods for which direct observations can be indispensable.[104]

Clearly label negative findings. A study that analyzed >200 articles in the field of Animal Cognition found large heterogeneity in how non-significant effects were labeled, including various instances of i) ambiguous and imprecise wording and ii) misinterpretation of non-significant results as support for the null hypothesis, the latter being prevalent in >80% in titles.[105] Similar effects were also found in other fields.[9,106] Therefore, if a non-significant result is obtained in a study, the finding should be clearly labeled, such as "The analysis did not show a significant effect of the manipulation" and effect sizes and/or confidence intervals should be discussed to increase transparency. Statements

such as "A was similar to B" or "There was no effect of A on B" are not justified based on non-significant results because a failure to reject H_0 does not confirm its correctness or generalizability. When reporting negative findings, one should not exclude the possibility of large individual variation, meaning that other individuals in a different setting might have provided a positive finding. Probably the most extreme example is the gray parrot (*Psittacus erithacus*) Alex. Although just one individual, much has been learned from him about numerical abilities and abstract abilities, thereby showing the potential capabilities of a species.[107]

CONCLUSIONS

Negative findings are valuable findings and fundamental to better understanding the world around us - they are part of all scientific discoveries. However, negative findings can be true or false, and one needs to be cautious to not confuse them, as false negative findings can lead to wrong conclusions. Here, we highlighted causes that can lead to false negative findings with the aim to start a discussion on how to distinguish true from false negative findings and when (and how) to report them. We hope that our recommendations will be helpful for designing rigorous studies that result in conclusive findings, independent of whether they are positive or negative.

AUTHOR CONTRIBUTIONS

MKS and JGF conceived the work; MKS wrote the first draft of the article which was expanded by JGF. MKS and JGF approved the final version of the article, which MKS submitted.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Sagan, C. (1977). The Dragons of Eden: Speculations on the Evolution of Human Intelligence. (Random House).

2. Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. Scientometrics 90, 891–904. https://doi.org/10.1007/s11192-011-0494-7.

3. Forstmeier, W., Wagenmakers, E.-J., and Parker, T.H. (2017). Detecting and avoiding likely false-positive findings: A practical guide. Biol. Rev. 92, 1941–1968. https://doi.org/10.1111/brv.12315.

4. Clark, A. (2017). Negative results: A crucial piece of the scientific puzzle. EveryONE. https://everyone.plos.org/2017/10/26/negative-results-a-crucial-piece-of-the-scientific-puzzle/.

5. Anon. (2019). The importance of no evidence. Nat. Hum. Behav. 3, 197. https://doi.org/10.1038/s41562-019-0569-7.

6. Journal of Negative Results in BioMedicine. https://jnrbm.biomedcentral.com/.

7. John, I. (2005). Why most published research findings are false. PLoS Med. 2, 696–701.

8. Hartgerink, C.H.J., Wicherts, J.M., and van Assen, M.A.L.M. (2017). Too good to be false: Nonsignificant results revisited. Collabra Psychol 3, 9. https://doi.org/10.1525/collabra.71.

9. Aczel, B., Palfi, B., Szollosi, A., Kovacs, M., Szaszi, B., Szecsi, P., Zrubka, M., Gronau, Q.F., van den Bergh, D., and Wagenmakers, E.-J. (2018). Quantifying support for the null hypothesis in psychology: An empirical investigation. Adv. Methods Pract. Psychol. Sci. 1, 357–366. https://doi.org/10.1177/2515245918773742.

10. Edelsbrunner, P.A., and Thurn, C.M. (2024). Improving the utility of non-significant results for educational research: A review and recommendations. Educ. Res. Rev. 42, 100590. https://doi.org/10.1016/j.edurev.2023.100590.

11. Cohen, J. (1992). Statistical power analysis. Curr. Dir. Psychol. Sci. 1, 98–101. https://doi.org/10.1111/1467-8721.ep10768783.

12. Farrar, B.G., Altschul, D.M., Fischer, J., van der Mescht, J., Placì, S., Troisi, C.A., Vernouillet, A., Clayton, N.S., and Ostojić, L. (2020). Trialling meta-research in comparative cognition: Claims and statistical inference in animal physical cognition. Anim. Behav. Cogn. 7, 419–444. https://doi.org/10.26451/abc.07.03.09.2020.

13. Milinski, M. (1997). How to avoid the seven deadly sins in the study of behavior. Adv. Study Behav. 26, 159–180. https://doi.org/10.1016/S0065-3454(08)60379-4.

14. Jennions, M.D., and Møller, A.P. (2003). A survey of the statistical power of research in behavioral ecology and animal behavior. Behav. Ecol. 14, 438–445. https://doi.org/10.1093/beheco/14.3.438.

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

16. Mehler, D.M.A., Edelsbrunner, P.A., and Matić, K. (2019). Appreciating the significance of non-significant findings in Psychology. J. Eur. Psychol. Stud. 10, 1–7. https://doi.org/10.5334/e2019a.

17. Farrar, B.G., Voudouris, K., and Clayton, N.S. (2021). Replications, comparisons, sampling and the problem of representativeness in animal behavior and cognition research. Anim. Behav. Cogn. 8, 273–295. https://doi.org/10.26451/abc.08.02.14.2021.

18. Webster, M.M., and Rutz, C. (2020). How STRANGE are your study animals? Nature 582, 337–340. https://doi.org/10.1038/d41586-020-01751-5.

19. Aguenounon, G., Aguenounon, G., Allritz, M., Altschul, D., Ballesta, S., Beaud, A., Bohn, M., Brandão, A., Brandão, A., Brooks, J., et al. (2020). The evolution of primate short-term memory. Anim. Behav. Cogn. 9, 428–516. https://doi.org/10.26451/abc.09.04.06.2022.

20. Reznick, D.N., and Ghalambor, C.K. (2005). Selection in nature: Experimental manipulations of natural populations. Integr. Comp. Biol. 45, 456–462. https://doi.org/10.1093/icb/45.3.456.

21. Siebeck, U.E. (2014). Communication in the ultraviolet: Unravelling the secret language of fish. In Biocommunication of Animals, G. Witzany, ed. (Springer Netherlands), pp. 299–320. https://doi.org/10.1007/978-94-007-7414-8_17.

22. Macías Garcia, C., and de Perera, T. (2002). Ultraviolet-based female preferences in a viviparous fish. Behav. Ecol. Sociobiol. 52, 1–6. https://doi.org/10.1007/s00265-002-0482-2.

23. Rick, I.P., and Bakker, T.C.M. (2008). Color signaling in conspicuous red sticklebacks: Do ultraviolet signals surpass others? BMC Evol. Biol. 8, 189. https://doi.org/10.1186/1471-2148-8-189.

24. Modarressie, R., Rick, I.P., and Bakker, T.C.M. (2006). UV matters in shoaling decisions. Proc. Biol. Sci. 273, 849–854. https://doi.org/10.1098/rspb.2005.3397.

25. Sabol, A.C., Hellmann, J.K., Gray, S.M., and Hamilton, I.M. (2017). The role of ultraviolet coloration in intrasexual interactions in a colonial fish. Anim. Behav. 131, 99–106. https://doi.org/10.1016/j.anbehav.2017.06.027.

26. Premack, D., and Woodruff, G. (1978). Does the chimpanzee have a theory of mind? Behav. Brain Sci. *1*, 515–526. https://doi.org/10.1017/S0140525X00076512.

27. Call, J., and Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. Trends Cogn. Sci. *12*, 187–192. https://doi.org/10.1016/j.tics.2008.02.010.

28. Hare, B., Call, J., and Tomasello, M. (2001). Do chimpanzees know what conspecifics know? Anim. Behav. *61*, 139–151. https://doi.org/10.1006/anbe.2000.1518.

29. Reimers, M., and Oakley, B. (2017). Empathy, theory of mind, cognition, morality, and altruism. In On Human Nature, M. Tibayrenc and F.J. Ayala, eds. (Academic Press), pp. 355–363. https://doi.org/10.1016/B978-0-12-420190-3.00021-1.

30. Krupenye, C., and Call, J. (2019). Theory of mind in animals: Current and future directions. Wiley Interdiscip. Rev. Cogn. Sci. *10*, e1503. https://doi.org/10.1002/wcs.1503.

31. Wright, D., and Krause, J. (2006). Repeated measures of shoaling tendency in zebrafish (*Danio rerio*) and other small teleost fishes. Nat. Protoc. *1*, 1828–1831. https://doi.org/10.1038/nprot.2006.287.

32. Schweinfurth, M.K., and Call, J. (2021). Capuchins (*Sapajus apella*) and their aversion to inequity. In Comparative Cognition, J. Anderson and H. Kuroshima, eds. (Springer), pp. 173–195. https://doi.org/10.1007/978-981-16-2028-7_11.

33. Barbano, M.F., and Cador, M. (2005). Various aspects of feeding behavior can be partially dissociated in the rat by the incentive properties of food and the physiological state. Behav. Neurosci. *119*, 1244–1253. https://doi.org/10.1037/0735-7044.119.5.1244.

34. Gross, J., Woelbert, E., Zimmermann, J., Okamoto-Barth, S., Riedl, A., and Goebel, R. (2014). Value signals in the prefrontal cortex predict individual preferences across reward categories. J. Neurosci. *34*, 7580–7586. https://doi.org/10.1523/JNEUROSCI.5082-13.2014.

35. Oustric, P., Thivel, D., Dalton, M., Beaulieu, K., Gibbons, C., Hopkins, M., Blundell, J., and Finlayson, G. (2020). Measuring food preference and reward: application and cross-cultural adaptation of the Leeds Food Preference Questionnaire in human experimental research. Food Qual. Prefer. *80*, 103824. https://doi.org/10.1016/j.foodqual.2019.103824.

36. Proserpio, C., Almli, V.L., Sandvik, P., Sandell, M., Methven, L., Wallner, M., Jilani, H., Zeinstra, G.G., Alfaro, B., and Laureati, M. (2020). Cross-national differences in child food neophobia: A comparison of five European countries. Food Qual. Prefer. *81*, 103861. https://doi.org/10.1016/j.foodqual.2019.103861.

37. Tennie, C., and Call, J. (2023). Unmotivated subjects cannot provide interpretable data and tasks with sensitive learning periods require appropriately aged subjects: A commentary on Koops et al. (2022) "Field experiments find no evidence that chimpanzee nut cracking can be independently innovated." Anim. Behav. Cogn. *10*, 89–94. https://doi.org/10.26451/abc.10.01.05.2023.

38. Brosnan, S.F. (2017). Understanding social decision-making from another species' perspective. Learn. & Behav. *46*, 101–102. https://doi.org/10.3758/s13420-017-0302-1.

39. House, B.R., Silk, J.B., Lambeth, S.P., and Schapiro, S.J. (2014). Task design influences prosociality in captive chimpanzees (*Pan troglodytes*). PLoS One *9*, e103422. https://doi.org/10.1371/journal.pone.0103422.

40. Jones, N.A.R., Cortese, D., Munson, A., Spence-Jones, H.C., Storm, Z., Killen, S.S., Bethel, R., Deacon, A.E., Webster, M.M., and Závorka, L. (2023). Maze design: Size and number of choices impact fish performance in cognitive assays. J. Fish. Biol. *103*, 974–984. https://doi.org/10.1111/jfb.15493.

41. Schweinfurth, M.K., and Call, J. (2019). Revisiting the possibility of reciprocal help in non-human primates. Neurosci. Biobehav. Rev. *104*, 73–86. https://doi.org/10.1016/j.neubiorev.2019.06.026.

42. Marshall-Pescini, S., Dale, R., Quervel-Chaumette, M., and Range, F. (2016). Critical issues in experimental studies of prosociality in non-human species. Anim. Cogn. *19*, 679–705. https://doi.org/10.1007/s10071-016-0973-6.

43. Hurst, J.L., and West, R.S. (2010). Taming anxiety in laboratory mice. Nat. Methods *7*, 825–826. https://doi.org/10.1038/nmeth.1500.

44. Brown, C. (2001). Familiarity with the test environment improves escape responses in the crimson spotted rainbowfish, *Melanotaenia duboulayi*. Anim. Cogn. *4*, 109–113. https://doi.org/10.1007/s100710100105.

45. Makaras, T., Stankevičiūtė, M., Šidagytė-Copilas, E., Virbickas, T., and Razumienè, J. (2021). Acclimation effect on fish behavioural characteristics: Determination of appropriate acclimation period for different species. J. Fish. Biol. *99*, 502–512. https://doi.org/10.1111/jfb.14740.

46. Jones, N.A.R., Newton-Youens, J., and Frommen, J.G. (2024). Rise and fall: Increasing temperatures have nonlinear effects on aggression in a tropical fish. Anim. Behav. *207*, 1–11. https://doi.org/10.1016/j.anbehav.2023.10.008.

47. Balzarini, V., Taborsky, M., Wanner, S., Koch, F., and Frommen, J.G. (2014). Mirror, mirror on the wall: The predictive value of mirror tests for measuring aggression in fish. Behav. Ecol. Sociobiol. *68*, 871–878. https://doi.org/10.1007/s00265-014-1698-7.

48. Chang, L., Fang, Q., Zhang, S., Poo, M.M., and Gong, N. (2015). Mirror-induced self-directed behaviors in rhesus monkeys after visual-somatosensory training. Curr. Biol. *25*, 212–217. https://doi.org/10.1016/j.cub.2014.11.016.

49. Kohda, M., Hotta, T., Takeyama, T., Awata, S., Tanaka, H., Asai, J.Y., and Jordan, A.L. (2019). If a fish can pass the mark test, what are the implications for consciousness and self-awareness testing in animals? PLoS Biol. *17*, e3000021. https://doi.org/10.1371/journal.pbio.3000021.

50. Krause, M.A., and Beran, M.J. (2020). Words matter: Reflections on language projects with chimpanzees and their implications. Am. J. Primatol. *82*, e23187. https://doi.org/10.1002/ajp.23187.

51. Cibulski, L., Wascher, C.A.F., Weiß, B.M., and Kotrschal, K. (2014). Familiarity with the experimenter influences the performance of common ravens (*Corvus corax*) and Carrion crows (*Corvus corone corone*) in cognitive tasks. Behav. Processes *103*, 129–137. https://doi.org/10.1016/j.beproc.2013.11.013.

52. Rabdeau, J., Badenhausser, I., Moreau, J., Bretagnolle, V., and Monceau, K. (2019). To change or not to change experimenters: Caveats for repeated behavioural and physiological measures in Montagu's harrier. J. Avian Biol. *50*, 1–12. https://doi.org/10.1111/jav.02160.

53. Krasheninnikova, A., and Schneider, J.M. (2014). Testing problem-solving capacities: Differences between individual testing and social group setting. Anim. Cogn. *17*, 1227–1232. https://doi.org/10.1007/s10071-014-0744-1.

54. Tomasello, M., and Call, J. (2011). Methodological challenges in the study of primate cognition. Science *334*, 1227–1228. https://doi.org/10.1126/science.1213443.

55. Whiten, A., Goodall, J., McGrew, W.C., Nishida, T., Reynolds, V., Sugiyama, Y., Tutin, C.E.G., Wrangham, R.W., and Boesch, C. (1999). Cultures in chimpanzees. Nature *399*, 682–685. https://doi.org/10.1038/21415.

56. Allan, A.T.L., Bailey, A.L., and Hill, R.A. (2020). Habituation is not neutral or equal: Individual differences in tolerance suggest an overlooked personality trait. Sci. Adv. *6*, eaaz0870. https://doi.org/10.1126/sciadv.aaz0870.

57. Altmann, S.A., and Altmann, J. (2003). The transformation of behaviour field studies. Anim. Behav. *65*, 413–423. https://doi.org/10.1006/anbe.2003.2115.

58. Baglione, V., Marcos, J.M., and Canestrari, D. (2002). Cooperatively breeding groups of carrion crow (*Corvus corone corone*) in Northern Spain. Auk *119*, 790–799. https://doi.org/10.1093/auk/119.3.790.

59. Bessa, J., Hockings, K., and Biro, D. (2021). First evidence of chimpanzee extractive tool use in Cantanhez, Guinea-Bissau: Cross-community variation in honey dipping. Front. Ecol. Evol. *9*, e625303. https://doi.org/10.3389/fevo.2021.625303.

60. Schradin, C., and Hayes, L.D. (2017). A synopsis of long-term field studies of mammals: achievements, future directions, and some advice. J. Mammal. *98*, 670–677. https://doi.org/10.1093/jmammal/gyx031.

61. Clutton-Brock, T., and Sheldon, B.C. (2010). Individuals and populations: The role of long-term, individual-based studies of animals in ecology and evolutionary biology. Trends Ecol. Evol. *25*, 562–573. https://doi.org/10.1016/j.tree.2010.08.002.

62. Soldati, A., Fedurek, P., Crockford, C., Adue, S., Akankwasa, J.W., Asiimwe, C., Asua, J., Atayo, G., Chandia, B., Freymann, E., et al. (2022). Dead-infant carrying by chimpanzee mothers in the Budongo Forest. Primates *63*, 497–508. https://doi.org/10.1007/s10329-022-00999-x.

63. Jack, K.M., Lenz, B.B., Healan, E., Rudman, S., Schoof, V.A.M., and Fedigan, L. (2008). The effects of observer presence on the behavior of *Cebus capucinus* in Costa Rica. Am. J. Primatol. *70*, 490–494. https://doi.org/10.1002/ajp.20512.

64. Carter, G.G., Schino, G., and Farine, D. (2019). Challenges in assessing the roles of nepotism and reciprocity in cooperation networks. Anim. Behav. *150*, 255–271. https://doi.org/10.1016/j.anbehav.2019.01.006.

65. Maher, C.R., and Lott, D.F. (2000). A review of ecological determinants of territoriality within vertebrate species. Am. Midl. Nat. *143*, 1–29. https://doi.org/10.1674/0003-0031(2000)143[0001:AROEDO]2.0.CO;2.

66. Dwan, K., Altman, D.G., Arnaiz, J.A., Bloom, J., Chan, A.W., Cronin, E., Decullier, E., Easterbrook, P.J., Von Elm, E., Gamble, C., et al. (2008). Systematic review of the empirical evidence of study publication bias and outcome reporting bias. PLoS One *3*, e3081. https://doi.org/10.1371/journal.pone.0003081.

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

68. van Assen, M.A.L.M., van Aert, R.C.M., Nuijten, M.B., and Wicherts, J.M. (2014). Why publishing everything is more effective than selective publishing of statistically significant results. PLoS One *9*, 84896. https://doi.org/10.1371/journal.pone.0084896.

69. Devine, S., Bautista-Perpinya, M., Delrue, V., Gaillard, S., Jorna, T., van der Meer, M., Millett, L., Pozzebon, C., and Visser, J. (2020). Science fails. J. Trial Error *1*, 1–5. https://doi.org/10.36850/ed1.

70. Gulinello, M., Mitchell, H.A., Chang, Q., Timothy O'Brien, W., Zhou, Z., Abel, T., Wang, L., Corbin, J.G., Veeraragavan, S., Samaco, R.C., et al. (2019). Rigor and reproducibility in rodent behavioral research. Neurobiol. Learn. Mem. *165*, 106780. https://doi.org/10.1016/j.nlm.2018.01.001.

71. Bohlen, M., Hayes, E.R., Bohlen, B., Bailoo, J.D., Crabbe, J.C., and Wahlsten, D. (2014). Experimenter effects on behavioral test scores of eight inbred mouse strains under the influence of ethanol. Behav. Brain Res. *272*, 46–54. https://doi.org/10.1016/j.bbr.2014.06.017.

72. Hampton, J.O., MacKenzie, D.I., and Forsyth, D.M. (2019). How many to sample? Statistical guidelines for monitoring animal welfare outcomes. PLoS One *14*, 0211417. https://doi.org/10.1371/journal.pone.0211417.

73. Nelson, L.D., Simmons, J.P., and Simonsohn, U. (2012). Let's publish fewer papers. Psychol. Inq. *23*, 291–293. https://doi.org/10.1080/1047840X.2012.705245.

74. Landhuis, E. (2016). Information overload. Nature *535*, 457–458. https://doi.org/10.1038/nj7612-457a.

75. Smaldino, P.E., and McElreath, R. (2016). The natural selection of bad science. R. Soc. Open Sci. *3*, 160384. https://doi.org/10.1098/rsos.160384.

76. Shephard, T.V., Lea, S.E.G., and Hempel de Ibarra, N. (2015). "The thieving magpie"? No evidence for attraction to shiny objects. Anim. Cogn. *18*, 393–397. https://doi.org/10.1007/s10071-014-0794-4.

77. Schmelz, M., Grueneisen, S., and Tomasello, M. (2020). The psychological mechanisms underlying reciprocal prosociality in chimpanzees (*Pan troglodytes*). J. Comp. Psychol. *134*, 149–157. https://doi.org/10.1037/com0000200.

78. Bailoo, J.D., Reichlin, T.S., and Würbel, H. (2014). Refinement of experimental design and conduct in laboratory animal research. ILAR J. *55*, 383–391. https://doi.org/10.1093/ilar/ilu037.

79. Dankel, S.J. (2019). What information is provided from non-significant findings and how can this be improved? J. Trainology *8*, 19–23. https://doi.org/10.17338/trainology.8.2_19.

80. Garamszegi, L.Z. (2016). A simple statistical guide for the analysis of behaviour when data are constrained due to practical or ethical reasons. Anim. Behav. *120*, 223–234. https://doi.org/10.1016/j.anbehav.2015.11.009.

81. Mara, C.A., and Cribbie, R.A. (2012). Paired-samples tests of equivalence. Commun. Stat. Simul. Comput. *41*, 1928–1943. https://doi.org/10.1080/03610918.2011.626545.

82. Lakens, D. (2017). Equivalence tests: a practical primer for t-tests, correlations, and meta-analyses. Soc. Psychol. Personal. Sci. *8*, 355–362. https://doi.org/10.1177/1948550617697177.

83. Lakens, D., Scheel, A.M., and Isager, P.M. (2018). Equivalence testing for psychological research: a tutorial. Adv. Methods Pract. Psychol. Sci. *1*, 259–269. https://doi.org/10.1177/2515245918770963.

84. Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. Front. Psychol. *5*, 781. https://doi.org/10.3389/fpsyg.2014.00781.

85. Konijn, E.A., van de Schoot, R., Winter, S.D., and Ferguson, C.J. (2015). Possible solution to publication bias through Bayesian statistics, including proper null hypothesis testing. Commun. Methods Meas. *9*, 280–302. https://doi.org/10.1080/19312458.2015.1096332.

86. Morey, R.D., and Rouder, J.N. (2011). Bayes factor approaches for testing interval null hypotheses. Psychol. Methods *16*, 406–419. https://doi.org/10.1037/a0024377.

87. Lee, S.Y., and Song, X.Y. (2004). Evaluation of the Bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. Multivar. Behav. Res. *39*, 653–686. https://doi.org/10.1207/s15327906mbr3904_4.

88. Henrich, J., Heine, S.J., and Norenzayan, A. (2010). Most people are not WEIRD. Nature *466*, 29. https://doi.org/10.1038/466029a.

89. Voelkl, B., and Würbel, H. (2016). Reproducibility crisis: Are we ignoring reaction norms? Trends Pharmacol. Sci. *37*, 509–510. https://doi.org/10.1016/j.tips.2016.05.003.

90. Jones, N.A.R., Webster, M.M., and Salvanes, A.G.V. (2021). Physical enrichment research for captive fish: Time to focus on the DETAILS. J. Fish. Biol. *99*, 704–725. https://doi.org/10.1111/jfb.14773.

91. Würbel, H. (2001). Ideal homes? Housing effects on rodent brain and behaviour. Trends Neurosci. *24*, 207–211. https://doi.org/10.1016/S0166-2236(00)01718-5.

92. Richter, S.H., Garner, J.P., and Würbel, H. (2009). Environmental standardization: Cure or cause of poor reproducibility in animal experiments? Nat. Methods *6*, 257–261. https://doi.org/10.1038/nmeth.1312.

93. Frank, M.C. (2015). The ManyBabies Project. https://manybabies.github.io/.

94. Altschul, D.M., Beran, M.J., Bohn, M., Call, J., DeTroy, S., Duguid, S.J., Egelkamp, C.L., Fichtel, C., Fischer, J., Flesset, M., et al. (2019). Establishing an infrastructure for collaboration in primate cognition research. PLoS One *14*, e0223675. https://doi.org/10.31234/osf.io/3xu7q.

95. Grandin, T. (1989). Behavioral principles of livestock handling. Prof. Anim. Sci. *5*, 1–11. https://doi.org/10.15232/S1080-7446(15)32304-4.

96. Krupenye, C., Kano, F., Hirata, S., Call, J., and Tomasello, M. (2016). Great apes anticipate that other individuals will act according to false beliefs. Science *354*, 110–114. https://doi.org/10.1126/science.aaf8110.

97. Pritchard, D.J., Hurly, T.A., Tello-Ramos, M.C., and Healy, S.D. (2016). Why study cognition in the wild (and how to test it)? J. Exp. Anal. Behav. *105*, 41–55. https://doi.org/10.1002/jeab.195.

98. Hare, B. (2001). Can competitive paradigms increase the validity of experiments on primate social cognition? Anim. Cogn. *4*, 269–280. https://doi.org/10.1007/s100710100084.

99. Janmaat, K.R.L. (2019). What animals do not do or fail to find: A novel observational approach for studying cognition in the wild. Evol. Anthropol. *28*, 303–320. https://doi.org/10.1002/evan.21794.

100. Crofoot, M.C., Lambert, T.D., Kays, R., and Wikelski, M.C. (2010). Does watching a monkey change its behaviour? Quantifying observer effects in habituated wild primates using automated radiotelemetry. Anim. Behav. *80*, 475–480. https://doi.org/10.1016/j.anbehav.2010.06.006.

101. Rutz, C., Bluff, L.A., Weir, A.A.S., and Kacelnik, A. (2007). Video cameras on wild birds. Science *318*, 765. https://doi.org/10.1126/science.1146788.

102. Gao, B., Wotton, K.R., Hawkes, W.L.S., Menz, M.H.M., Reynolds, D.R., Zhai, B.P., Hu, G., and Chapman, J.W. (2020). Adaptive strategies of high-flying migratory hoverflies in response to wind currents: Flight behaviour of migrant hoverflies. Proc. Biol. Sci. *287*, 20200406. https://doi.org/10.1098/rspb.2020.0406rspb20200406.

103. Dunlop, K.M., Jarvis, T., Benoit-Bird, K.J., Waluk, C.M., Caress, D.W., Thomas, H., and Smith, K.L. (2018). Detection and characterisation of deep-sea benthopelagic animals from an autonomous underwater vehicle with a multibeam echosounder: A proof of concept and description of data-processing methods. Deep-Sea Res. Part A Oceanogr. Res. Pap. *134*, 64–79. https://doi.org/10.1016/j.dsr.2018.01.006.

104. Smith, J.E., and Pinter-Wollman, N. (2021). Observing the unwatchable: Integrating automated sensing, naturalistic observations and animal social network analysis in the age of big data. J. Anim. Ecol. *90*, 62–75. https://doi.org/10.1111/1365-2656.13362.

105. Farrar, B.G., Vernouillet, A., Garcia-Pelegrin, E., Legg, E.W., Brecht, K.F., Lambert, P.J., Elsherif, M., Francis, S., O'Neill, L., Clayton, N.S., and Ostojić, L. (2023). Reporting and interpreting non-significant results in animal cognition research. PeerJ *11*, e14963. https://doi.org/10.7717/peerj.14963.

106. Fidler, F., Burgman, M.A., Cumming, G., Buttrose, R., and Thomason, N. (2006). Impact of criticism of null-hypothesis significance testing on statistical reporting practices in conservation biology. Conserv. Biol. *20*, 1539–1544. https://doi.org/10.1111/j.1523-1739.2006.00525.x.

107. Pepperberg, I.M. (2006). Grey parrot numerical competence: A review. Anim. Cogn. *9*, 377–391. https://doi.org/10.1007/s10071-006-0034-7.