Acta Neuropsychiatrica

cambridge.org/neu

Tips and traps for behavioural animal experimentation

Shokouh Arjmand[1], Gregers Wegener[1,2], Anne M. Landau[1,3] and Amanda Eskelund[1]

[1]Translational Neuropsychiatry Unit, Department of Clinical Medicine, Aarhus University, Aarhus, Denmark; [2]AUGUST Center, Department of Clinical Medicine, Aarhus University, Aarhus, Denmark and [3]Department of Nuclear Medicine and PET, Department of Clinical Medicine, Aarhus University and Hospital, Aarhus, Denmark

Perspective

Cite this article: Arjmand S, Wegener G, Landau AM, and Eskelund A. (2022) Tips and traps for behavioural animal experimentation. *Acta Neuropsychiatrica* **34**:240–252. doi: 10.1017/neu.2022.4

Received: 7 September 2021
Revised: 14 January 2022
Accepted: 24 January 2022
First published online: 3 February 2022

Key words:
animal experimentation; research design; research techniques; biostatistics; behaviour; animal

Author for correspondence:
Amanda Eskelund, Email: ares@clin.au.dk

Abstract

Behavioural animal experimentation is an inseparable part of research trying to understand the biological underpinnings of human behaviour, diseases and disorders. Working with animals comes with great responsibility to achieve reliable and reproducible results of highest scientific quality. In a simple step-by-step fashion, we highlight some common issues that may occur along the path to conducting behavioural animal experimentations and posit some solutions and grounds to ensure the excellence of work done in this research area while aspiring to improve conditions for laboratory animals. It entails topics of study design, animal and experimenter welfare, experimental considerations and frequentist biostatistics. At the end, we direct to some guidelines and manuals that may prove valuable to researchers in this field. Our ten simple tips and traps are meant for students who are learning about important concepts for the first time; graduates whose statistics training all too often has neglected the concept of power in experimental design; and researches who would like a light-hearted refresher on these topics. With this perspective, we hope that you will avoid falling into traps and find answers to what you always wanted to know about conducting behavioural animal experimentation.

Summation

- In this perspective, you will be introduced to some conspicuous ups and downs that you may experience while doing behavioural animal experimentation.
- Through a ten-step footpath, we attempt to give you a standpoint on what could be of potential importance, and how to render your best behavioural animal experiment while providing most optimal conditions for the laboratory animals.

Perspectives

- We are still far from abandoning use of animals; therefore, we must strive to optimise and refine animal welfare, while continuously improving the quality of our experiments to advance translatability of animal models and replicability of results.
- Conducting an ideal behavioural animal experimentation does not only lean on the animals but also the experimenters. Know yourself very well and be prepared.
- Having a good knowledge of biostatistics could be a key to successful behavioural animal experimentation.

Introduction

Look into the mirror. What stares back at you? A reflection of a human silhouette or something beyond that, which is more personified? For centuries humans have been preoccupied with the question 'who am I?', a seemingly simple question we are still grappling to define. Our identity and personality are hugely shaped by our behaviours and we have some evidence that our brain is involved in governing these behaviours. But questions still remain unanswered. What has shaped your behaviour? What controls it? And what will happen if these controls are broken? Can they be fixed?

As human beings, we are naturally drawn to studies of the brain and how it is involved in our decision-making processes and behaviours since these are vital parts of our daily lives. However, getting to grips with the intricacy of how the brain works to shape behaviours is also essential for improving human health and well-being. Causal manipulations of the brain are important for the characterisation of its functions, yet invasive methods have only limited applicability in human research due to ethical concerns.

The importance of animals in research is highlighted by the fact that animals have been a vital part of work done by the majority of Nobel laureates in medicine or physiology, ranging from Ivan Pavlov in 1904 (the father of classical conditioning) to Tinbergen, von Frisch and Lorenz in 1973 (founders of ethology) and more recently to Moser, Moser and O'Keefe in 2014 (neurobiology of spatial navigation). Through such works, behavioural animal experimentation has provided a strong foundation for generating knowledge that is considered to be of great benefit to all humankind and has shifted our view of scientific paradigms.

Studies in comparative or translational behavioural neuroscience encompass studying an animal's response, interaction and adaptation inside an experimental assay while aiming to elucidate the underlying biological mechanisms. This embraces both modelling human symptoms and conditions in animals as well as studying innate behaviours because we interpret distinct behaviours as evolutionarily adaptive traits.

Working with animals comes with great responsibility and requires respect for the animals as living beings. International and national ethics commissions have implemented legal directives to protect animal use in science and researchers must adhere to the 3R strategy where animals are only used in an experiment if there is no alternative. Therefore, researchers do their best to substitute research on animals if possible and try to use alternative investigative methods where applicable (*Replacement*); attempt to improve experimental conditions for animals and mitigate any pain, distress or discomfort that might be due to the experiment (*Refinement*); and finally endeavour to decrease the number of animals used in the study or gain more data from the same number of animals (*Reduction*). Researchers have an important moral responsibility to ensure highest standard of animal welfare – not only due to legal requirements but also because the welfare of our animals influences the quality of our research.

The quality of animal experimentation is of utmost importance to the researcher and is limited by the degree of translatability to the human condition. Some fields of animal experimentation have suffered a bad reputation in terms of generating information that is poorly translated to humans, and therefore, doubt has been cast on the merit and utility of this field (Pound *et al.*, 2004). Main issues include poor efficacy of pharmaceuticals in humans after encouraging animal model data or lack of ability to replicate animal studies (O'Collins *et al.*, 2006; Belzung, 2014; Garner, 2014). Suggested solutions to the issues raised encourage researchers to have a higher awareness of all modalities in their experiments, covering study design, animal management, reporting and improved statistical practices (Ioannidis, 2005; Garner, 2014; Aarts *et al.*, 2015). Ultimately, when working with animals we have to ensure that our studies are conducted with consideration for *all* possible aspects of the study, while also weighing the ethical elements.

Our goal is to advance the quality and replicability of work done in the field while aspiring to improve conditions for laboratory animals and the experimenter. Because the merit of our research will be evaluated by our statistical analyses, it is an advantage to ensure to design and conduct the studies so the basis for drawing high-quality interferences is in place. For example, be aware that there are always four possible outcomes to your upcoming study (Fig. 1). Therefore, we have collected and highlighted some important tips and traps at every step of a pathway to conduct animal experimentation (see map in Fig. 2).

All roads may lead to Rome, but not all roads are safe. Our goal is to guide the reader on a path to conducting high-quality research, travelling through neighbourhoods of biostatistics, quality data and animal welfare. On our suggested path we have ten stops along the way, where we preinform the traveller of some of the traps and suggest some helpful tips to ease the journey. We have designed it for students learning concepts for the first time, graduate students and researchers venturing into new research areas, and those who would like a light-hearted refresher. Our topics are broad to target the audience in behavioural animal experimentation, but some in-depth descriptions and discussions are suggested at the end. The route and the ten stops we suggest are conservative, it may not be the most direct path, and many alternative routes exist; however, it is a path that is well-travelled and well-illuminated. Let's get going.

Stop 1: Start by defining focused hypotheses, for your own sake!

Our first stop includes the tip to 'start off with a focused hypothesis'. Focused hypotheses are the torches that light up your road toward answering the questions on which your study will be built. Generating hypotheses upfront and based on a solid theoretical framework will dramatically help you to plan ahead. A hypothesis can be focused (often used in a confirmatory study) or it can be broad 'are there any differences?' (hypothesis-generating study).

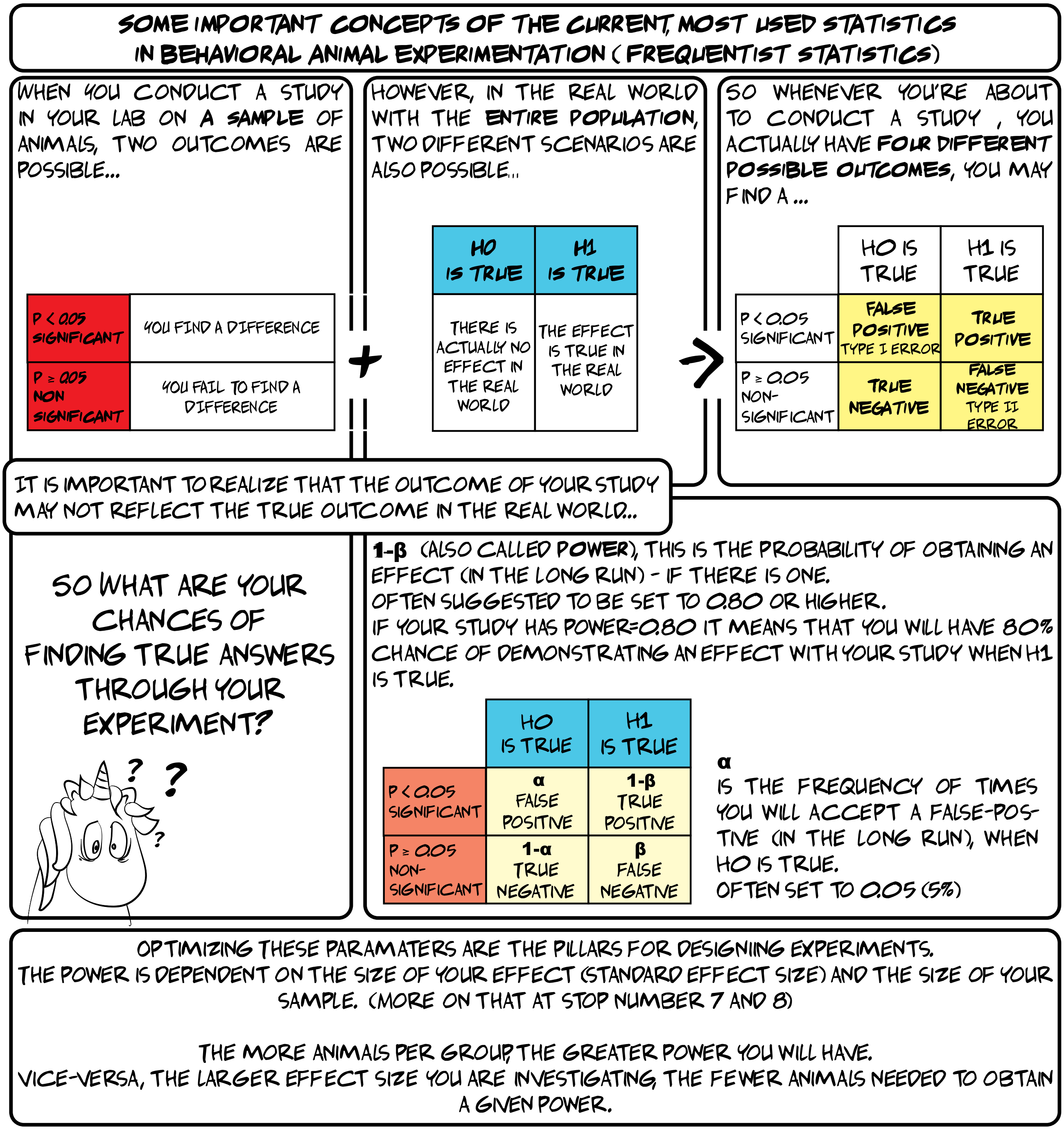
Sticking to a simple, focused hypothesis instead of a broad one will enable you to estimate your sample size effectively, develop contingency plans, design each step of your research project in great detail and collect data more responsibly. Additionally, simpler, confirmatory studies have the advantage that they can help establish scientific findings as replicable.

Hypothesis-generating studies do not start off with a focused hypothesis and often employ a 'shot-gun' approach, investigating multiple, diverse targets simultaneously. The main drawback to hypothesis-generating studies is that they lack the statistical power to provide confidence in the results, they could however be followed up by confirmatory studies. The advantage of hypothesis-generating studies is that they are important for gaining information on variables of interest and model parameters, prior to designing a confirmatory study. In that sense, hypothesis-generating studies are intimately linked to hypothesis confirming studies (Biesecker, 2013). Even in a hypothesis-driven study, it is perfectly acceptable to explore and report secondary findings as long as it is explicitly noted that they are indeed exploratory and their statistical limitations are discussed (more about this at stop 8). Therefore, a study can be based on a primary hypothesis and report explorative secondary findings.

Avoid 'hypothesizing after the results are known' (HARKing), which means you formulate your hypothesis after you have seen the results. HARKing is a malpractice that may lead to theories based on false findings (e.g. type I errors as mentioned in Fig. 1 (Kerr, 1998; Brian Haynes, 2006)). Once published, such false-positive findings may prove hard to eradicate and waste enormous amounts of research resources. Therefore, the scientific costs of HARKing probably outweigh its benefits, and additionally, it cannot be of help to you if you should find yourself feeling lost at any time *while* you are conducting the animal study. HARKing can be avoided by pre-registering your study (Biesecker, 2013) (as suggested at stop 9), where you demonstrate that your hypotheses were indeed determined prior to data collection.

If your study is taking unexpected turns and you have lost your way, your hypothesis acts as your beacon. It reminds you why you are doing the study and the importance of finishing it – regardless of whether your hypothesis later on proves to be right or wrong.

https://doi.org/10.1017/neu.2022.4 Published online by Cambridge University Press



SOME IMPORTANT CONCEPTS OF THE CURRENT, MOST USED STATISTICS IN BEHAVIORAL ANIMAL EXPERIMENTATION (FREQUENTIST STATISTICS)

WHEN YOU CONDUCT A STUDY IN YOUR LAB ON A **SAMPLE** OF ANIMALS, TWO OUTCOMES ARE POSSIBLE...

P < 0.05 SIGNIFICANT	YOU FIND A DIFFERENCE
P ≥ 0.05 NON SIGNIFICANT	YOU FAIL TO FIND A DIFFERENCE

+

HOWEVER, IN THE REAL WORLD WITH THE **ENTIRE POPULATION**, TWO DIFFERENT SCENARIOS ARE ALSO POSSIBLE...

H0 IS TRUE	H1 IS TRUE
THERE IS ACTUALLY NO EFFECT IN THE REAL WORLD	THE EFFECT IS TRUE IN THE REAL WORLD

→

SO WHENEVER YOU'RE ABOUT TO CONDUCT A STUDY , YOU ACTUALLY HAVE **FOUR DIFFERENT POSSIBLE OUTCOMES**, YOU MAY FIND A ...

	HO IS TRUE	H1 IS TRUE
P < 0.05 SIGNIFICANT	FALSE POSITIVE TYPE I ERROR	TRUE POSITIVE
P ≥ 0.05 NON-SIGNIFICANT	TRUE NEGATIVE	FALSE NEGATIVE TYPE II ERROR

IT IS IMPORTANT TO REALIZE THAT THE OUTCOME OF YOUR STUDY MAY NOT REFLECT THE TRUE OUTCOME IN THE REAL WORLD...

SO WHAT ARE YOUR CHANCES OF FINDING TRUE ANSWERS THROUGH YOUR EXPERIMENT?

1-β (ALSO CALLED **POWER**), THIS IS THE PROBABILITY OF OBTAINING AN EFFECT (IN THE LONG RUN) - IF THERE IS ONE.
OFTEN SUGGESTED TO BE SET TO 0.80 OR HIGHER.
IF YOUR STUDY HAS POWER=0.80 IT MEANS THAT YOU WILL HAVE 80% CHANCE OF DEMONSTRATING AN EFFECT WITH YOUR STUDY WHEN H1 IS TRUE.

	HO IS TRUE	H1 IS TRUE
P < 0.05 SIGNIFICANT	α FALSE POSITIVE	1-β TRUE POSITIVE
P ≥ 0.05 NON-SIGNIFICANT	1-α TRUE NEGATIVE	β FALSE NEGATIVE

α
IS THE FREQUENCY OF TIMES YOU WILL ACCEPT A FALSE-POSITIVE (IN THE LONG RUN), WHEN HO IS TRUE.
OFTEN SET TO 0.05 (5%)

OPTIMIZING THESE PARAMATERS ARE THE PILLARS FOR DESIGNIING EXPERIMENTS.
THE POWER IS DEPENDENT ON THE SIZE OF YOUR EFFECT (STANDARD EFFECT SIZE) AND THE SIZE OF YOUR SAMPLE. (MORE ON THAT AT STOP NUMBER 7 AND 8)

THE MORE ANIMALS PER GROUP, THE GREATER POWER YOU WILL HAVE.
VICE-VERSA, THE LARGER EFFECT SIZE YOU ARE INVESTIGATING, THE FEWER ANIMALS NEEDED TO OBTAIN A GIVEN POWER.

Fig. 1. Four possible outcomes to a study.

This simple tip might be undervalued, but it is a fundamental prerequisite to continue our path to the upcoming stops 2, 3 and 4. So, always keep in mind that it is perfectly fine if your results do not support your initial hypotheses; that is why the hypotheses are called '*a priori*' and need to be rigorously tested through experimentation.

Stop 2: Should animals be used to assess the hypothesis?

This stop is most relevant for students and early-career researchers, but it is nevertheless a prudent exercise to continuously reflect on this question. We stopped at 'define your hypothesis' before 'should you use animals?', simply because once you have established your hypothesis, you will have an easier time deciding whether a behavioural animal experiment is even suitable for its investigation and whether you should pursue this research field.

If you have already established that animals are necessary to study your hypothesis let's move on to the fact that ethics is very important to researchers in animal experimentation. Scientists are Human! Humans are filled with passion, empathy, emotion and ambition. They *do* care and that's why they have chosen to put the majority of their lives into exploring trajectories for a better

https://doi.org/10.1017/neu.2022.4 Published online by Cambridge University Press

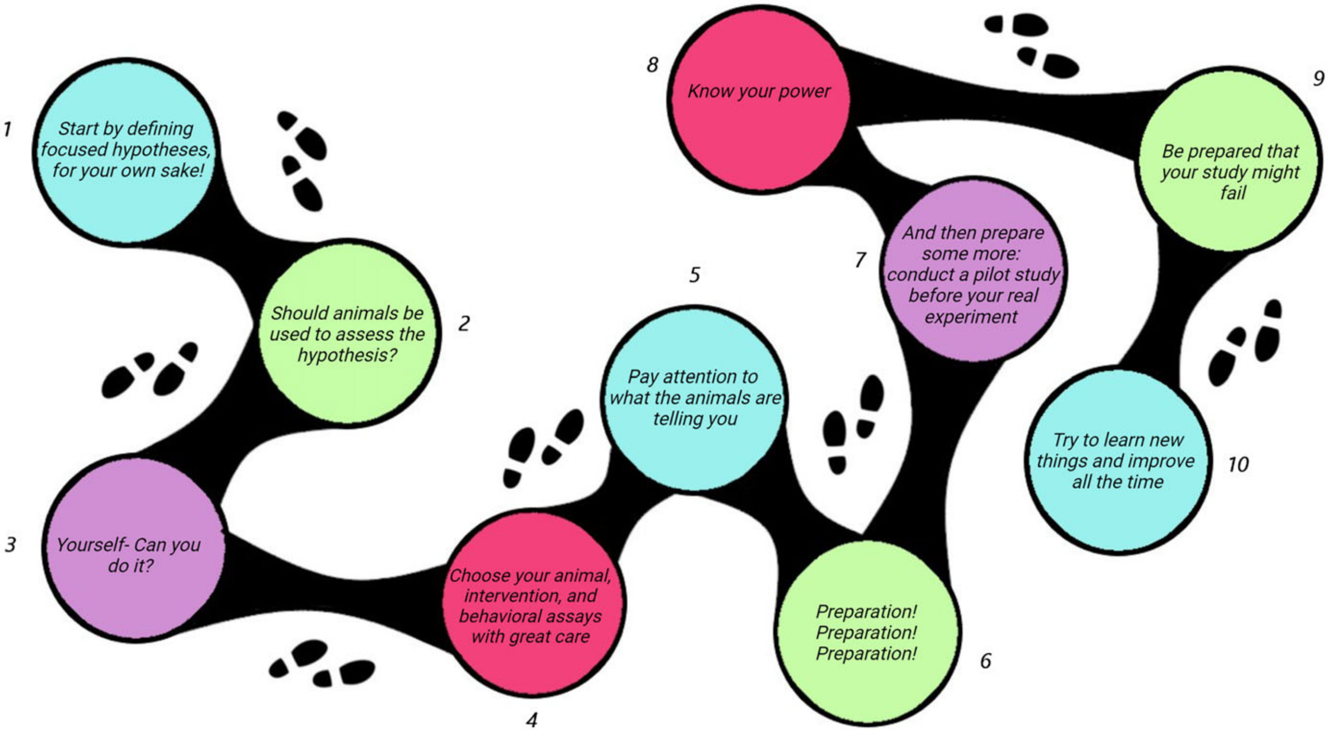

1
Start by defining focused hypotheses, for your own sake!
2
Should animals be used to assess the hypothesis?
3
Yourself- Can you do it?
4
Choose your animal, intervention, and behavioral assays with great care
5
Pay attention to what the animals are telling you
6
Preparation! Preparation! Preparation!
7
And then prepare some more: conduct a pilot study before your real experiment
8
Know your power
9
Be prepared that your study might fail
10
Try to learn new things and improve all the time

Fig. 2. Our suggested route and stops to behavioural animal experimentation.

life and pushing the human race forward. Nevertheless, since experimental animals may experience stress, pain and discomfort, scientists are constantly faced with the dilemma of striving to find an intervention to improve human well-being, but with the cost to the living animal. How much pain, suffering and distress of experimental animals should be allowed when conducting research? Fortunately, frameworks to which you can always refer when planning experiments were constructed more than 50 years ago, such as the 3Rs. So here is another fact; scientists do their best to replace research on animals whenever possible and use animals *only* if there is no alternative. Thus, they ensure to reduce the number of animals used in a study and refine the methods to alleviate the discomfort to the animals (Percie du Sert *et al.*, 2020). Also, animal studies go through a harm-benefit analysis, so only animal studies that generate valuable new information are approved.

Animals are called for in several types of studies. We still do not have adequate knowledge of all the interactions, components and pathways involved in computing a symptom or behaviour, so we are still far from *in silico*. Additionally, cells, organs and biological systems are surprisingly interconnected; thus, cell culture studies (though applicable and constructive) can hardly justify the superfluity of animal research. In vivo behavioural studies sustain the very elaborate interconnectivity of different organs. Furthermore, with animals, you have the ability to explore biological phenomena concurrently, parallel to or even after observing behaviour in a controlled milieu. This makes behavioural animal studies the kernel for the transference of in vitro/*ex* vivo and in vivo research to human applications (Homberg *et al.*, 2021).

If experiments on living laboratory animals are indispensable and cannot be substituted to clarify your hypothesis, carrying out the experiment in the best possible way is a must. As the next step, you need to assess yourself (next stop, number 3) and search literature that explores and describes translatability and back-translatability (how humans behave in the animal assays) of your chosen animal model (at stop number 4). So, let's move on.

Stop 3: Yourself – can you do it?

Before reaching the point of no return on your path to embarking on a scientific career in animal behaviour, it will be wise to weigh both the pros and cons of this line of work. Assess early on whether your personality fits well with this line of research and whether you have 'got what it takes' to conduct and complete the work with animals.

Patience is an important skill you need to have. Expect to have to wait for your protocol to be ethically approved, which can take an unpredictable amount of time, so you may end up postponing experiments. Additionally, behavioural animal experiments are generally very time-consuming and require a lot of patience, diligence and the stamina of a horse. You may spend hours and hours handling your animals to habituate them to you and conduct extensive training that requires daily sessions for an extended period of time (yes, also on holidays and weekends!). You may work alone in a cellar without windows as behavioural tests are sensitive to the presence of other people or environmental factors (see under stop 6). The circadian rhythm of your animal may be different from yours, which can involve working during non-working hours. Often this work is not conducted in ergonomically comfortable positions and you may also feel the physical load of your work at the end of the day. Don't let this discourage you, but just be aware that this is part of the job under some circumstances and that it is usually time-limited.

This work also requires that you will spend a large portion of your time in direct contact with the animals, and considering your own health is a top priority. Allergies are very unpleasant, can develop spontaneously and may gradually worsen over time. Masks are often used for the prevention or alleviation of allergy symptoms but may be uncomfortable or impractical when interacting with animals for longer periods. Check and follow the safety procedures in your laboratory. Clear guidelines should be available and consider whether you will want to engage in work of for example, zoonotic diseases and other risk hazards if you are pregnant or immunodeficient.

Our tip is that the work with animals will be a lot more rewarding if you have a genuine interest in interacting with the animals and concern for their well-being. In general, this type of work has several advantages; you work away from a typical office setting during the periods of your experiment and two days are rarely the same. Although the animals are not your pets, it is very satisfying to follow their progress, for example, to see them gradually become more comfortable around you. Maybe they can even be hand-fed small rewards or they automatically search for you when you approach them because they have learned you feed them snacks (you are nice!). Since your success is partly dependent on the success of the animals (they are indeed your most valued collaborators!), it is very rewarding when this joint partnership comes to fruition. For example, in the form of animals demonstrating they have successfully mastered a training paradigm.

One very crucial step to consider is whether you have the ability to end an animal experiment yourself and try it hands-on. You may have spent hours with each animal, and it may seem cruel to end an experiment. However, with your hypothesis in place (stop 1), and clear reasoning as to why it is important to use animals in your study (stop 2), this should make it more tolerable to you. This work is never going to get easy. Even though your experienced colleagues can make it look cushy, don't make the mistake of believing it is an undemanding chore for them. It is rarely a sustainable solution to make others do this part of the experiment for you if you are planning to conduct behavioural studies as a career. We believe that being able to finish the work you have started allows you to better preserve a touch of reality and use it, for example, to reflect harder about the necessity of a study before you initiate it. Learning how to do it is indeed a gradual process, that falls under some of the immense preparation work involved in conducting this type of study (covered at stop number 6).

If your answer is yes to the question raised in the above headline and you meet all the above-mentioned criteria, it is time to jump to our next stop.

Stop 4: Choose your animal, intervention, and behavioural assays with great care

It is super important to keep in mind that this tip represents the majority of the planning stage of your study and bad choices may cause irreparable detriments to your behavioural experiment. Hence, gathering adequate knowledge of models, interventions, and assays is imperative to design a rigorous study and you should expect to spend a lot of time studying before starting a study. This part of your study design is highly field-dependent and we suggest finding relevant systematic reviews and meta-analyses and discussing them thoroughly with your supervisor(s) and co-workers. Some of the overall topics to pay attention to are described below.

Choosing an appropriate model or animal can be a game-changer and carefully consider choice of species, strain, and sex, and whether to use induced models (e.g., pharmacological, mechanical or environmental), spontaneous models, transgenic models, negative models, or orphan models (Poindron *et al.*, 2008). Based on your formulated hypotheses you should choose your animal model wisely and expect to do a lot of research on it. It is highly recommended to always consider whether the animal model you have chosen is the best to put your hypotheses to test (Box 1). If you conduct translational neuroscience, remember that a single animal model rarely represents the whole disease or disorder. Instead, an animal model could be considered a model to study certain pathology or symptoms of a disease.

Get to know any drugs and vehicles you are working with like the back of your hand, knowing the physicochemical properties of your experimental compound can help you do your study excellently. With drugs that are tricky to be dissolved, think long and hard about what you can do to ensure successful administration and achieve meaningful and ethical data. Check whether other formulations of your drug are available (e.g. clinically approved formulations), or whether it is better to replace the drug or use another administration route (more examples in Box 1).

Being informed and critical will also help you when you decide on your experimental assay. When conducting several tests in a combined behavioural battery design, remember that previous experiment(s) might have a protracted impact and in some way affect the outcome of the following assays, so always consider carry-over effects (Blokland *et al.*, 2012) (See examples in Box 1). It is generally advised to perform the most stressful test as the last test and rest animals in between tests (Crawley, 2007). Try to be as knowledgeable as possible regarding the tests you are about to carry out and challenge yourself as to whether you can improve anything. Be aware that if you plan to later study the brain itself, a stressful behaviour test may confound your findings.

Also, be aware that one experimental assay may have a completely different paradigm when another animal model or intervention is used than the one originally intended (see example in Box 1). It is strongly advised to keep the choice of animal model, intervention, and assay procedures constant and standardised to avoid variations in results (more about this at stop number 6). The method section in scientific journals is rarely very comprehensive, so if your goal is to replicate work from another lab, don't hesitate to contact the authors right off the bat and ask about every detail on an animal model, intervention and assay during your preparation.

Basically, to generate valuable knowledge it is essential to design your experiments so the animal model, intervention, and assay come together in the most meaningful way. In a nutshell, you have to be aware of the nuts and bolts of every step of your experiment.

Stop 5: Pay attention to what the animals are telling you

Remember that your animals did not enter your study with informed consent. Animals cannot speak for themselves directly, and they cannot leave the study whenever they want to. Hence, treat them *respectfully*, handle them *gently* and make an effort to try to *understand* their behaviour and beware of what that behaviour is telling you.

During the period where you introduce yourself to the animals, why not give them a thorough physical check-up? Basically, you act as their doctor and systematically check for gross abnormalities in both sensory and motor function, and make sure animals appear to be healthy (bodyweight, fur coat and state of teeth can be checked daily!). Find great guidelines with suggestions on how to assess the general health of laboratory rodents in

https://doi.org/10.1017/neu.2022.4 Published online by Cambridge University Press

Box 1. Some important factors to consider when planning your study

Model/Animal

– If you work with a chemically induced lesion in the brain that is representative of the final stages of a neurodegenerative disease, could it tell you anything about the early stages and the aetiology of the disease?
– If a specific strain of an animal is more resistant/sensitive to stress, which one should be used to induce a model representing human stress-induced disorder, for example early life adversity?
– Drug metabolism is sex-dependent; so can you generalise the findings of a dose-response study performed only on one sex of animals to the target population? (external validity)
– Consider the ethology of your model: for example, the choice of using nocturnal or diurnal animals to model mood disorders.

Intervention

– Some drugs cannot be administered via a special route of administration and you should always consider the bioavailability if you want to convert between iv., ip., icv., sc., or po.
– Knowing your drug's storage condition and stability, its physicochemical properties (such as photosensitivity, stability, viscosity, solubility, pH, colour, rheology), as well as its pharmacology, potential toxicity and pharmacokinetics is essential for ensuring meaningful administration to an animal without causing adverse effects such as irritation or pain.
– For injections – do you use fresh syringes and needles for each animal (and accept the loss of drug in dead space) or do you reuse (limited amount of drug, accept the risk of carry-over unknowns that could cause inflammation)
– Some compounds might be of safety concerns to yourself (e.g. carcinogenic, corrosive) and need special care while working with them or disposing leftover material, so always read the datasheet before you start working with new drugs and compounds.

Experimental Assay

– Always consider whether a combination of assays may provide basis for carry-over effect in a behavioural battery.
Consider whether combining two in an unfortunate way can provide basis for false positives.
An example is the combination of two commonly used tests of depression-like behaviour, for example sweet-tasting gustatory preference test (anhedonia: consumes less sweets) and forced swim test (despair: floats more). If animals are first given cookies (in a test of anhedonia) and then shortly after placed in the forced swim test (test of behavioural despair), the group of animals that consumed fewer cookies (the anhedonic ones), could have less energy and therefore float more in the swim test, simply because of their lower caloric intake, and not due to 'despair'. This principle also applies to the opposite test order; consider whether animals that were more active and spent more energy in a swim test, may eat more cookies if they are given the opportunity right after swimming. The caloric intake and energy metabolism therefore could become confounding factors of the results.
– With experimental designs that use 'between-groups' comparisons (e.g. animals receive either treatment or vehicle, never both): consider whether it is possible to obtain a baseline of behaviour or biological specimen before the actual intervention is conducted, to ensure treatment effects only became apparent after the treatment (i.e. are not due to unfortunate randomisation).

Combinations

– Consider whether an assay paradigm changes if you switch species, sex, or age. For example, the natural social structure differs for rats and mice. Mice are suggested to be more aggressive towards conspecifics than rats (Ellenbroek and Youn, 2016). This could affect the interpretation of social behaviour assays.
– Gustatory discriminatory tests are typically used to assess anhedonia. However, if you work with a model with an impaired sense of taste or intervention that alters the senses involved in taste, this assay may instead have become a criterion of successful manipulation of the animal model rather than a test of anhedonia. Similarly, a drug may have adverse effects such as polydipsia (i.e. induce excessive thirst), which also affects any readout based on drinking solutions.
– Locomotor-based tests (like forced swim test) need to be well-analysed to avoid any misinterpretation. If you do experiments that infer changes in energy metabolism and compare groups that have different muscular mass/strength, body fat content, sex, and age, etc. remember to interpret results with this caveat in mind.

the literature (Crawley, 2007). An example to imagine how relevant this is could be Pavlov's auditory-conditioned training of dogs. Dogs were trained to associate the sound of a bell with food, once fully trained the sound alone would make the dogs drool. Now imagine if a dog was deaf to begin with, that dog would have to be eliminated from the study after several months of hard work. By assessing the general health of the animals beforehand (such as their hearing or vision), you can eliminate the risk of conducting a long-term study only to realise later that some animals had an abnormality, became outliers and had to be excluded from your data. Having to exclude animals after data collection unfortunately also makes you lose power in your experiment (more about that at our 8th stop). Therefore, by making the effort of assessing the general health of your animals before you conduct an experiment, you can avoid the use of animals in vain and save time.

Keep an eye on the animals' well-being and welfare continuously to ensure that you abide by the humane endpoints. Checking your animals daily also allows you to keep track of whether animals always have adequate water, chow and whether their ventilation system is working. This may seem a bit trivial, but even a short-duration change in either of these vital factors

Box 2. Examples of causes of variation

Examples of **biological variation** include differences in:

- age
- sex
- strain (background strain and number of back-crossings)
- species
- colony (vendor or in-house breeding)
- hormones (oestrous cycle and corticosterone level)

Experimental variation includes, but is not limited to, changes in:

- handling/no handling and which experimenter(s) interact with the animals
- diet, water
- housing (number of animals, enrichment, etc.)
- experimental conditions (temperature, humidity, light, etc.)
- acclimatisation period to facility and rooms before the experiment
- time of day, or time of year (due to hormones and other rhythmic activities in the facility of which you may (not) be aware of)
- variation in equipment used for behavioural assay
- the method chosen to clean equipment between animals
- noise, smell/scents, or other potential disturbances from adjacent rooms
- details about the specimen, its preparation, and storage, for example, transplantation or surgery

Sources of **variations in outcome variable**:

- tracking of behaviour: different manual methods or automated
- scoring behaviour: different manual methods or automated
- dichotomising continuous variables and differences in binning (e.g. splitting animals into responders/non-responders based on a mean or median. For a discussion on disadvantages on such procedures, see (Lazic, 2018)).

Make sure to report such factors (see the ARRIVE guidelines) in your publication to permit replicability

may dramatically alter your results. So you would want to know all the details, first-hand, if an unfortunate event like that ever happened!

Once you have started your study, note down any unspecific effects. Preclinical studies often investigate novel drugs, of which little is known. For example, you may notice the paws of treated animals are colder than those of animals not treated with this drug, or that a novel mouse strain always makes right-sided turns. Although the consequence of this finding may be unknown at the time of discovery, it is a good idea to report it for later analysis and comparison.

Stop 6: Preparation! Preparation! Preparation!

It is mandatory to take courses on laboratory animal sciences before you start working with live animals. However, you may have to learn behavioural assays that are specific to your workplace. For this purpose, it is so important to develop a very detailed, step-by-step protocol and include all materials needed. To get you started, consult other official guidelines, for example the PREPARE guidelines developed by Norecopa (Smith *et al.*, 2018), which consists of a short checklist and accompanying webpage to help you think of things ahead of your experiment. Don't forget to also look at the ARRIVE guidelines, that were developed to put emphasis on reporting in animal experimental design, as this also helps during the planning stage (du Sert *et al.*, 2020).

When you are writing your own protocol, try to write it as complete as possible and think of everything ahead. Visualise how you are going to execute the experiment in advance and put in as many details as you can think of. The better you prepare, the readier you are if any unexpected incidents occur. If you are in doubt about anything, do ask your peers or supervisors. Ask someone with relevant experience to go over your protocol at least once for a critical appraisal.

However, knowing a behavioural test in theory is not enough, and you have to get involved practically to understand any caveats that may come up. Next, you do hands-on training before the commencement of an actual study. Co-op with a peer you trust and learn their methods to handle animals, conduct the experimental assay, and other practical issues of relevance (such as training a specific route of administration, step-wise learning how to finish an experiment). If you are setting up a new assay in your lab, peer-reviewed scientific video journals are extremely valuable (see Box 5). Develop the habit of jotting everything down while you receive training and occasionally pause in between to check up if you are on the right track and perform sanity checks. Ask for help whenever needed and whenever you think that it may improve the outcome of your study – no questions are too stupid!

The preparation part is also where you consider all possible instances that can influence your behavioural experiment and ensure quality data. Always consider circumstances from the animals' perspectives; they probably have a different sense of vision, hearing, smell and tactile stimuli, and therefore, things that seem negligible to you may have a profound impact on the animals. Take a close look at Box 2, which lists some important, commonly reported causes of variations in studies. Be aware that they might vary slightly from field to field within animal experimentation, but the important take-home message here is to **try to keep all factors consistent every time you repeat an experiment**. This is simply because unintentional changes in these small details may eventually lead you to reject your study *post hoc* if you didn't succeed in replicating prior results (more about that at stop number 9).

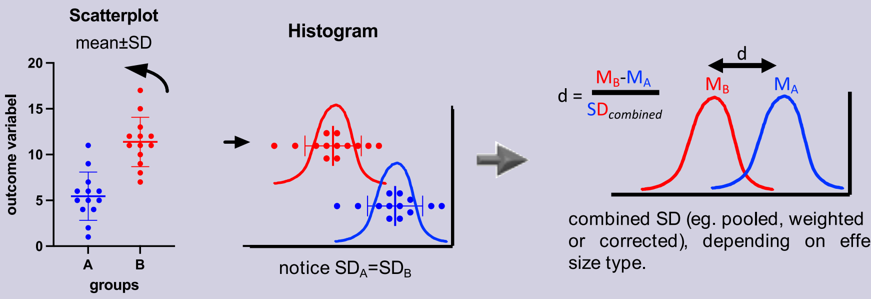
Because you want to ensure that your experimental conditions are consistent every time you conduct the experiment, there are some golden tips you must follow every time you enter a behavioural facility: Many animals have stronger senses than humans, so keep noise level way down and don't make conversations with colleagues, unless you are sure no one is running experiments. Avoid emitting strong odours, such as perfumes or garlic, as these may linger in the room long after you leave and potentially disturb a colleagues' experiment. Make sure both your experimental equipment and the room are clean when you start any experiment and also sweep the room and empty the garbage bins! The last things you want inside your room are unintended odour cues (e.g. poop or tissues used to wipe away urine), which could come from different species, strains, or sexes. Therefore, if several people are running experiments at the same time, schedule timeslots with each other so that no one disturbs each other's experiment unintentionally. Also, consider how you clean equipment between animals. Alcohol may eliminate odour cues but if ventilation is not efficient, it may be a very overwhelming experience to animals and cause them to respond differently.

Box 3. Effect sizes and *p*-values

Effect size can be unstandardised (e.g. the mean difference between two groups, keeps the variable unit) or standardised (unitless). *Standardised effect sizes* have many names for example, Cohen's d, correlational constant r, eta partial squared η_p^2, etc. As an example, Cohen's d is the number of standard deviations two means are apart from each other and shown below for two independent groups:

Your data may be plotted as a scatter plot with means and standard deviation (SD). Flip it 90 degrees to visualize it as histograms:

Cohen's d (d) is the two means (M) substracted from each other, divided by the pooled standard deviation (SD)

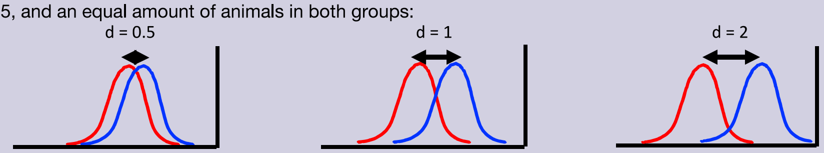


$$d = \frac{M_B - M_A}{SD_{combined}}$$

combined SD (eg. pooled, weighted or corrected), depending on effect size type.

Cohen's d can have both positive and negative values (this indicates the direction of the effect compared to the control), the further away from 0 the stronger the effect, whereas close to 0 means any effect is more likely to be unimportant. Because Cohen's d can be biased with small sample sizes, Hedge's g is an unbiased alternative that uses a weighted SD instead of a pooled SD, but the principle is similar. (Read more about some of the common effect sizes, how they are calculated and how they compare to each other here (Lakens, 2013).)

Lets look at some examples of what the effect size means in animal research. Three different values of Cohen's d, would require different 'least amount' of animals for an unpaired, 2-tailed t-test with α = 0.05, and an equal amount of animals in both groups:



To demonstrate d = 0.5	To demonstrate d = 1	To demonstrate d = 2
80% power ≥ 64 animals/group	80% power; n ≥ 17 /group	80% power; n ≥ 6 /group
95% power ≥ 105 animals/group	95% power; n ≥ 27/group	95% power; n ≥ 8/group

Standardised effect size gives you a number that tells you how big of a difference you found between your groups. The size of the p-value is not useful for this purpose, for example, you might be excited to get a really, really low p-value like $p = 0.0000000001$. However, you almost never know if you operate under a reality where H0 or H1 is true (as shown in Fig. 1), and your very low p-value could be a false-positive finding. In fact, if H0 is true in the real world, $p = 0.0000000001$ is just as probable as $p = 0.049$ or $p = 0.88$. (See the distribution of *P*-values when H0 is true in the Carlin cartoon (Fig. 3).) Thus, a lower p-value does not necessarily mean that you had a stronger effect, and in that sense, p-values can fool you.

As described in all the official guidelines (PREPARE, ARRIVE) always (always!) randomise your animals to both treatments and the order in which you test the animals (e.g. either completely randomised or randomised block design (Festing, 2020)). This is a crucial measure to avoid bias and confounding factors, such as circadian rhythm and biological/experimental variabilities. There may be factors in your study design that you expect will affect the variation, but that you are not interested in studying, such as body weight, age, litter, locomotor activity. Here it may be an advantage to stratify your groups/treatments within blocks according to these variables (for visual explanation, see randomised block design in (Festing, 2020) or Roslin Institute videos on randomisation techniques (Box 5)).Other options are to include them as cofactors in the statistical analysis or make sure to clearly state that the results are interpreted with this caveat in mind.

To become a good, unbiased observer, always keep an open mind and conduct all stages of your study as blinded as possible. This means, try to be blinded to experimental groups and treatments (if you can), so that you will remain unaware of which animals had what treatments until your data analysis is completely finished. Remember that nature has its own rules and things are not necessarily the way you expect them to be. So, keep searching for evidence with an unbiased mind and observe meticulously, then you may discover true and novel findings.

Finally, only start when you are confident that you can elegantly perform a behavioural animal study without inflicting stress or other unintentional discomforts on the animals due to being a novice. The next step is to carry out a pilot study of the experiment you want to do.

Stop 7: And then prepare some more: conduct a pilot study before your real experiment!

Experiments done on a low scale at first can simulate conditions of your real experiment, provide you with a first-hand idea of the parameters in your experiment and the effect size (See Box 3 for

https://doi.org/10.1017/neu.2022.4 Published online by Cambridge University Press

Box 4. Helpful tips for sample size calculations

Many statistics programmes can perform sample size calculations. An example is G * Power, which is a free programme that estimates sample size based on standardised effect size, alpha, and power.

Perform the sample size calculation according to the test you plan to use

– Standard effect sizes are convertible. Basically, if you have calculated partial eta squared, or omega squared from a one-way ANOVA, you can convert it into, for example, Cohen's d to calculate sample size for a t-test and vice-versa (instructions on how to calculate effects sizes and use them in G * Power (Lakens, 2013). For two-way ANOVA interactions, consider what type of interaction you expect and whether it is practically feasible to power correctly (Giner-Sorella, 2018).

Estimate the effect size of your primary aim

– For sample size calculations, you calculate based on the *minimum effect size* you are interested in detecting. If you have no clue of the effect of what you are about to investigate, using Cohen's famous benchmarks could be an option, for example, 'strong effect': $r > 0.5$ or Cohen's $d > 0.8$, but note that such effects sizes are low in animal research, which operates on homogenous animals from strictly controlled environments. It can be based on practical significance, theoretical predictions, or limitation of resources, but you have probably conducted one or more pilot studies that can help you narrow down the effect size. If you're lucky, a publication or meta-analysis can help. Just remember that published results may be subjected to p-hacking and publication bias, and you may have to adjust the effect size slightly closer to 0 to account for this. There will usually be uncertainties about what the true effect size is, so this is a difficult step.

Define your desired alpha and power level

– You may want higher power for experiments that are hard to conduct, to make sure you don't miss out on an effect if there really is one. If you are doing multiple comparisons, adjusting alpha accordingly could be relevant.

Adjust your sample size accordingly to the attrition rate

– Problems ensuring proper drug delivery, surgeries, inability to learn a required training programme, etc. are some examples of factors that may lead to dropouts of animals before your experiment is finalised. Make sure to supplement the calculated sample size if dropouts are expected.

Keep it simple

– As listed in stop 1, we suggest to keep a focused hypothesis and power your study towards your primary aim. Because you power your study according to this specific effect, your study will have high power to detect the interesting effect, low power for other effects. Since the 3R recommends 'getting as much information per animal as possible', it is not uncommon to see study designs where researchers look for several interactions. Interactions are where you investigate if your variable is dependent on several factors. For example, whether a drug (vs. vehicle) affects gene '*Supermouse*' differently in a disease model versus (vs. control mouse). This is a 'drug×animal_model' factorial design, where the aim is to find the interaction (e.g. the treatment worked differently in the disease model than the control model). You could potentially also be interested in whether this gene '*Supermouse'* was differently regulated in one brain region versus another and you would have a three-way interaction 'drug×animal_model×brain_region'. However, every time you add an interaction, you risk a loss of power, especially with ordinal interactions (Giner-Sorella, 2018). By focusing on one main aim with sufficient power level you're in a good position to confirm/reject your main hypothesis, and you can still report all other findings as explorative and discuss their limitations.

Optional stopping can be p-hacking

– An example of optional stopping is: A researcher notices a trend in one study and decides to repeat the study over and over, pooling subjects from all experiments until the trend reaches statistical significance. Instead of performing a study with an *a priori* determined sample size, the sample size was determined on results from interim analyses and pooled as if it all came from one experiment. Statistically, this malpractice will eventually lead to significant findings just by chance, and optional stopping inflates false-positive findings if not properly adjusted for.

definition of effect size) and you will always learn a lot from them. Therefore, it is essential to carry out a pilot study before you conduct your actual experiment.

Your pilot study could often be a replication or extension of either a method or model, derived from the literature or some work already carried out in your laboratory. Different terms have been coined to describe 'types of replication'; for example, direct replication (exact or similar study) or quasi (conceptual) replication (where one switches out the species or tries a different assay to demonstrate the same paradigm). For your first experiment, it is a good idea to start off with as direct replication as possible and leave conceptual replication for later on.

When you are ready to analyse your pilot data, remember to pay particular attention to any potential deviant-behaving animals *before* you unblind yourself and look at the data. Once you have plotted your pilot data, do not remove outliers, don't use an outlier test and don't remove data just because they are further away from the mean. This is a pilot study and you are going to repeat it with proper power and sample size. Your pilot study is intended to provide you with invaluable information on how your animals truly

https://doi.org/10.1017/neu.2022.4 Published online by Cambridge University Press

Box 5. Helpful links and resources

1. Check *guidelines* already when planning your study such as PREPARE (Smith *et al*., 2018) and https://norecopa.no/PREPARE
2. Checklist for when you are ready to report data ARRIVE (du Sert *et al*., 2020) and https://arriveguidelines.org/
3. Find recommendations on *how to report* your data with this guideline (Michel *et al*., 2020)
4. The *book* 'What's wrong with my laboratory mouse?' is good at every step of planning and conducting animal experiments with both rats and mice (Crawley, 2007)
5. *Videos* that explain concepts of animal study designs are found on YouTube from The Roslin Institute – Training, find channel 'Introduction to Experimental Design' (The Roslin Institue - Training, 2016)
6. A free tool to *randomise* anything – for example your list of animals: www.random.org/lists
7. *Pre-register* animal studies:
   - The Animal Study Registry (Bert *et al*., 2019) (https://www.animalstudyregistry.org/asr_web/index.action,
   - CAMARADES Systematic review and meta-analysis on preclinical studies, http://syrf.org.uk/
   - Center for Open Sciences pre-register (https://www.cos.io/initiatives/prereg)
8. *Peer-reviewed scientific video journals*:
   - JoVe (https://www.jove.com), more recent open-access video journals are being established (e.g. http://www.videojournalofbiomedicalscience.com/).
9. *Help on statistics*:
   - Daniël Lackens free course on statistical inference and common misunderstandings on the *p*-value (for beginners and experienced alike): https://www.coursera.org/learn/statistical-inferences/home/welcome
   - His excel templates for calculating different effect sizes Calculating_Effect_Sizes.xlsx; from_R2D2.xlsx for converting between effect size families and accompanying article (Lakens, 2013) can be found on the open science platform: https://osf.io/vbdah/.
10. Blog post that explains considerations for powering an interaction in a two-way ANOVA experiment (Giner-Sorella, 2018).
11. *Discussions* on how to increase power while reducing animal use:
    - Frequentist approaches, simple tips that can be included in any animal study design and may help to increase power (Lazic, 2018).
    - Bayesian approach, omit using control animals and rely on historical data (Bonapersona *et al*., 2021).
    - Suggestion of best method when conducting multi-batch studies (though not behaviour) (Karp *et al*., 2020).
    - Conducting animal studies more like human studies, where variation is preserved (Garner, 2014).

behave. In your next (complete) study, there will likely also be animals that behave differently from the mean. You desperately need that raw effect size that you just generated because it will take this inherent variation into account when you calculate the sample size needed for your main study. By removing outliers after you are unblinded to your pilot data, you will only make it harder for yourself to replicate your own study.

You can compare the standard effect sizes every time you do a study and replication (like a forest plot), and it will give you a clue as to whether your effect is reliably detected, something that p-values don't offer. Because they are standardised units, they are useful to compare the size of an effect between your studies or for comparison to other labs. An example of an unstandardised effect size is the percent difference in outcome with treatment versus vehicle, which may be more difficult to use for comparison between experiments and labs. That is why you mainly find standardised effect sizes in meta-analyses. Also, standardised effect sizes are a core part of doing an *a priori* sample size analysis for your real study.

Once this is settled, let's go to stop 8 and find out why this is so critical.

Stop 8: Know your power

Why is this important? The statistical **power** of an experiment is your probability of obtaining a significant result (for that specific effect) when your hypothesis is indeed correct. If you designed a study with a power of 80%, this means that on average, you *only* have an 80% chance of finding an effect ('a significant p-value') – even though there actually is one (e.g. a *true positive*, see Fig. 1). Yes! What this means is that even though a biological phenomenon exists (even though H1 is true), your ability to demonstrate it in a study, is not a given. This is because there will always be random noise in your data; after all, your sample is only a fraction of the entire population in the universe and there will probably be some imprecisions in your method. The power of your study depends on the size of the difference you look for with variation (e.g. effect size), alpha level and the number of animals per group. The higher your sample size, the greater power you will obtain. The greater the difference (effect size), the fewer animals you will need to obtain a given power.

The generic recommended minimum power of an experiment is 80%, but the higher it is, the better your chances of demonstrating an effect. It is important to realise, that even with 80% power, you have a 20% chance of missing out on finding a significant effect that is present (this is called a type II error, see Fig. 1). Therefore, the lower the power, the higher the chance of a type II error. Unfortunately, recent meta-analyses show an abyssal median power in published studies within neuroscience of 21% (Button *et al*., 2013) and less than 50% in studies of animal behaviour (Jennions and Møller, 2003). For two commonly used cognitive-behavioural rodent tests (Morris water maze and radial arm maze), it was 18% and 31%, respectively (Button *et al*., 2013). 21% power basically means there is only a 1:5 chance of demonstrating a phenomenon: On average, you would have to redo the entire experiment five times to demonstrate one significant result, even though there indeed is an effect (H1 is true). This also means that in four out of five studies; time, money, labour, and animals are wasted on studies that are unable to detect the true underlying effect.

The problems with conducting low-powered studies also entail other issues. The effect size obtained from them will be less precise (wider distribution) and often inflated (which means we risk overestimating the effectiveness of our treatment or model) (Button *et al.*, 2013). Significant results from under-powered studies are more likely to be a false-positive than a true positive (Christley, 2010). But even worse, low power tends to be propagated across future studies because future studies are based on the effect sizes from published studies (Ioannidis, 2005; Button *et al.*, 2013). Therefore, performing studies with low power have very little informational value (they are basically just noise) which raises serious ethical concerns in animal studies.

Hopefully, you will see why a thorough understanding of the concepts of power, sample size calculation and effect size must become a key prerequisite for animal research. Not all departments have a statistician available, but decisions on whether or not to initiate a study need to be made on a very frequent basis; therefore, it might be unrealistic to have to pay for outside statistical counselling. Some advice on the procedure and factors worth considering are given in Box 4 and we recommend taking a course, if you are not already familiar with performing *a priori* sample size calculations (See Box 5 for our suggestion to a free online Coursera course in English).

Now that you see why this simple tip is so crucial, start obtaining a greater signal-to-noise ratio in your hypothesis testing and avoid type II errors.

Stop 9: Be prepared that your study might fail

Once you have completed your pilot study, made a sample size calculation on appropriate power for your real experiment and decided on methods, it is ideal to pre-register your real study. If you are worried about competitors snatching your ideas, be relieved to find that pre-registering can be done without disclosing your detailed research plans for up to 5 years on The Animal Study Registry (Bert *et al.*, 2019). Note that this has to be done *before* you start collecting any data from subjects in your study. Hopefully, your real study goes as planned and you see your effect repeated in a properly designed study. However, we think it is a piece of good advice to warn you upfront about the possibility that your study might fail and why.

Recall that since the power of your study is most likely below 100%, you might not get the expected result purely because of variation in your data and methodological imprecisions. Replicating a published study, or conducting your own study twice, with both having 80% power, there is only a $0.8 * 0.8 = 0.64 = 64\%$ chance that you get a significant result in both studies. Remember that this applies even if your hypothesis is correct (H1 is true) and both experiments were conducted completely as intended. Statistics is a tricky business, so it might bring peace of mind to remember to account for this little fact before jumping to conclusions that 'this experiment didn't work because of X or Y (Box 2)', or rejecting your hypothesis prematurely. Use the effect sizes generated in all of your different studies to try to narrow down the true effect size for your next study.

If your study had strong power to detect your specific effect (where the type II error is unlikely) and enough experiments have been conducted to convince you that you are not trying to replicate a false-positive finding (the type I error is unlikely), then it may be relevant to dig into your data and explore whether some biological or experimental variability could account for the inability to replicate. This includes everything from animal strains, colonies, sex, age, time of day, noise and smells present during the experiment as already mentioned in Box 2. This can reveal intriguing facts about your model, intervention or assay, that may revolutionise the field so these are very important to also take into account. Remember though your study was powered to a specific effect and may have low power to detect other effects. Therefore, when you try to analyse for spurious findings *post hoc*, you are performing explorative and probably under-powered hypothesis-generating analyses. Before you can establish that these are in fact the reasons for the inability to replicate, a study may be designed with the appropriate power to specifically test these new hypotheses.

With these tips in mind, you may be thinking: 'why go through the bother of replicating if I can't even be sure that I will get the same result?' The fact is, to start finding out whether our significant results are true positives or false positives, we should replicate entire studies and openly share both the significant and non-significant studies to figure out the distribution of p-values. As illustrated in Fig. 3, replicating well-powered studies gives us an indication of whether our p-value distribution belongs to a reality where H0 is the truth or H1 is the truth, so we need replications to gain verisimilitude.

Another idea is to supplement frequentist statistics with Bayesian statistics, where you continuously update your prior hypothesis with your gathered data, which can be a particular advantage when replicating studies (Dienes, 2014; Bonapersona *et al.*, 2021). Unlike frequentist statistics, Bayesian statistics give you an estimate of how strong your evidence is in favour of a hypothesis, and it is also the only of the two forms of statistics that provide evidence for the H0 hypothesis. Remember that $p \geq 0.05$ in frequentist statistics does *not* mean 'there is no difference', it simply means you do not know whether there is a difference (your study could be under-powered).

Now that you know that you might not end up meeting your expectations it is super important to learn it was indeed anticipated to happen somewhere along the road. An important thing is to realise how important your study is regardless of the outcome. So, bounce right back on track and keep up the good work.

Stop 10: Try to learn new things and improve all the time

In a field that is incessantly being refined, we as researchers also have to continuously brush up our skills and gain new ones. This includes courses on new lab techniques that are relevant for your behavioural assay, establishing collaborations and getting to know new behavioural assays, etc. Consider to pair up with someone competent in your lab that you would really like to learn more from; ask them if you can participate or observe a study or watch videos to learn their method.

Less obvious skills to continuously brush up on throughout your scientific career include data presentation, statistics and drawing relevant inferences. These are the tools with which you cut your huge chunk of data into manageable-sized bites. Brushing up and learning new statistics is what keeps that knife sharp and increases your skill in how to make the most sensible cuts in your data.

Always break free of the 'we always did it that way'; this is an unfortunate practice that probably hinders scientific progress. Be a pioneer and try to figure things out even if no one else in your own lab knows. Reach out to other scientists to try to find your answer. Always stay critical and don't be scared of questioning and

https://doi.org/10.1017/neu.2022.4 Published online by Cambridge University Press

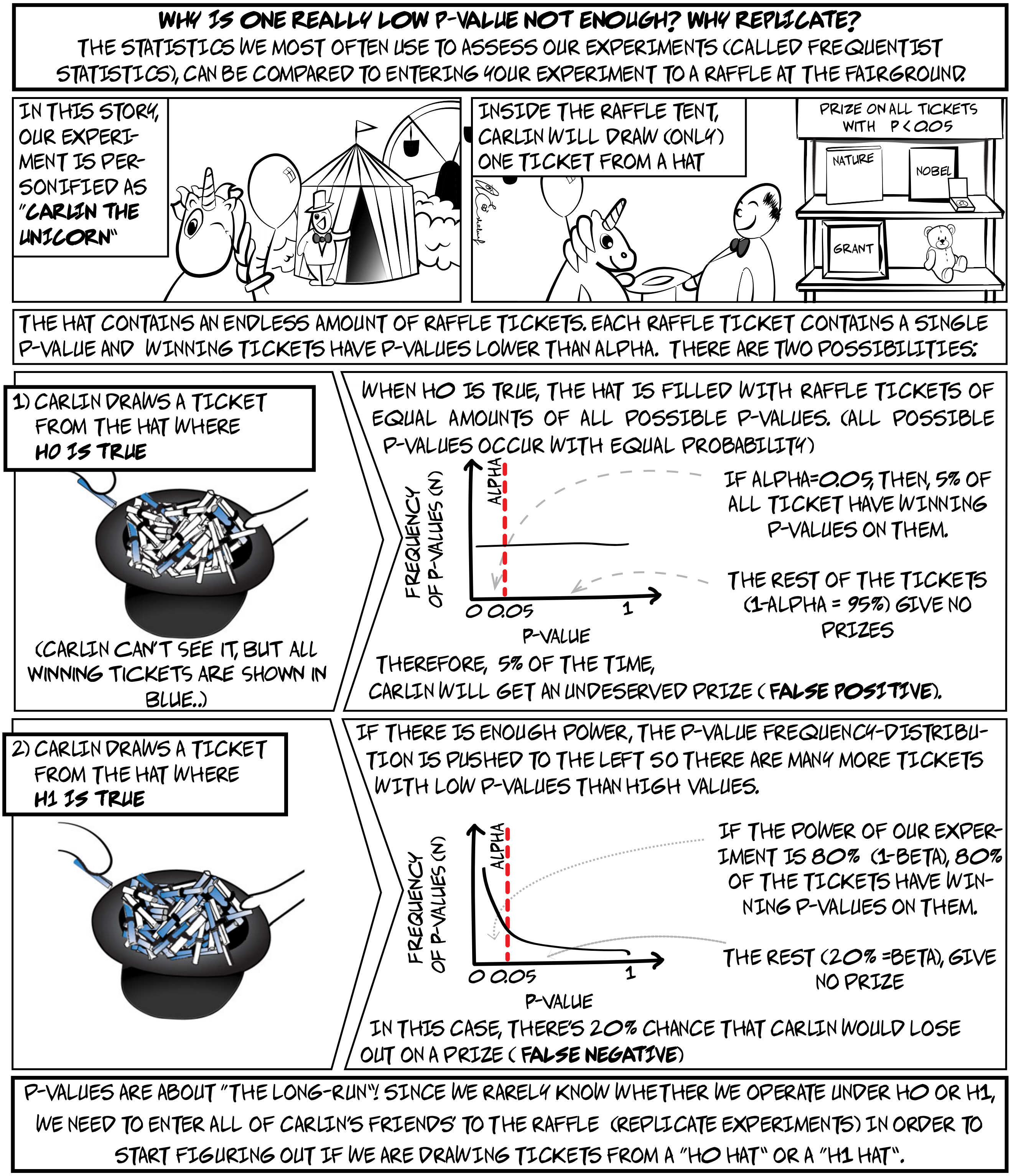

WHY IS ONE REALLY LOW P-VALUE NOT ENOUGH? WHY REPLICATE?
THE STATISTICS WE MOST OFTEN USE TO ASSESS OUR EXPERIMENTS (CALLED FREQUENTIST STATISTICS), CAN BE COMPARED TO ENTERING YOUR EXPERIMENT TO A RAFFLE AT THE FAIRGROUND.
IN THIS STORY, OUR EXPERIMENT IS PERSONIFIED AS "CARLIN THE UNICORN"
INSIDE THE RAFFLE TENT, CARLIN WILL DRAW (ONLY) ONE TICKET FROM A HAT
PRIZE ON ALL TICKETS WITH P<0.05
NATURE
NOBEL
GRANT
THE HAT CONTAINS AN ENDLESS AMOUNT OF RAFFLE TICKETS. EACH RAFFLE TICKET CONTAINS A SINGLE P-VALUE AND WINNING TICKETS HAVE P-VALUES LOWER THAN ALPHA. THERE ARE TWO POSSIBILITIES:
1) CARLIN DRAWS A TICKET FROM THE HAT WHERE H0 IS TRUE
(CARLIN CAN'T SEE IT, BUT ALL WINNING TICKETS ARE SHOWN IN BLUE..)
WHEN H0 IS TRUE, THE HAT IS FILLED WITH RAFFLE TICKETS OF EQUAL AMOUNTS OF ALL POSSIBLE P-VALUES. (ALL POSSIBLE P-VALUES OCCUR WITH EQUAL PROBABILITY)
FREQUENCY OF P-VALUES (N)
ALPHA
0 0.05
1
P-VALUE
IF ALPHA=0.05, THEN, 5% OF ALL TICKET HAVE WINNING P-VALUES ON THEM.
THE REST OF THE TICKETS (1-ALPHA = 95%) GIVE NO PRIZES
THEREFORE, 5% OF THE TIME, CARLIN WILL GET AN UNDESERVED PRIZE (FALSE POSITIVE).
2) CARLIN DRAWS A TICKET FROM THE HAT WHERE H1 IS TRUE
IF THERE IS ENOUGH POWER, THE P-VALUE FREQUENCY-DISTRIBUTION IS PUSHED TO THE LEFT SO THERE ARE MANY MORE TICKETS WITH LOW P-VALUES THAN HIGH VALUES.
FREQUENCY OF P-VALUES (N)
ALPHA
0 0.05
1
P-VALUE
IF THE POWER OF OUR EXPERIMENT IS 80% (1-BETA), 80% OF THE TICKETS HAVE WINNING P-VALUES ON THEM.
THE REST (20% =BETA), GIVE NO PRIZE
IN THIS CASE, THERE'S 20% CHANCE THAT CARLIN WOULD LOSE OUT ON A PRIZE (FALSE NEGATIVE)
P-VALUES ARE ABOUT "THE LONG-RUN"! SINCE WE RARELY KNOW WHETHER WE OPERATE UNDER H0 OR H1, WE NEED TO ENTER ALL OF CARLIN'S FRIENDS' TO THE RAFFLE (REPLICATE EXPERIMENTS) IN ORDER TO START FIGURING OUT IF WE ARE DRAWING TICKETS FROM A "H0 HAT" OR A "H1 HAT".

Fig. 3. Carlin and p-value distributions.

revisiting the existing contemporary methods and theories. Try to adapt your behavioural assays regularly, to better adhere to the 3Rs (refine, replace, reduce), thereby you may improve the way in which a specific behavioural test has always been performed in your laboratory. Sometimes, tasks can be designed more simply or there might be an alternative way of doing a test waiting to be implemented.

As always 'keep an open mind' and remember that there is always room for improvement. You can always learn new things from new people, new places or new sources.

Conclusion

While undertaking animal experimentation work, it is a must to strive to provide better conditions for laboratory animals and buckle down to deliver the highest quality animal experiments possible. All experimental designs come with trade-offs, and it is a skill to conduct animal experimentation while ensuring the maximal likelihood of success, keeping standards for scientific rigour high, accepting practical limits and safeguarding ethics; a skill that continuously develops throughout your career.

As discussed at our stop number 8, our field is currently struggling with a high amount of low-powered studies, which generate little trustworthy information. We have to improve our statistical processes and turn this around. Our guided map includes good statistical conduct in behavioural animal experimentation on similar grounds as ethics, quality experimental design and animal welfare, and we hope your journey in our simple, ten-step footpath provides inspiration. As you slowly develop and walk your own footpath (over and over), always remember to integrate the good habits, strive to increase scientific rigour and benefit the overall quality of our field.

'We hope that breaking behavioural animal experimentation into ten simple steps makes it easier for you to disentangle the mysteries behind behaviours as research investigators in neuroscience. Maybe the next time you look into a mirror, you will have a more lucid reflection of yourself'.

Acknowledgements. Before submitting this manuscript, it was discussed in the Translational Neuropsychiatry Unit's journal club and the authors are grateful to all Ph.D. students and PIs at the Translational Neuropsychiatry Unit who provided us with their invaluable feedback and comments on the manuscript. Special thanks to Simone Larsen Bærentzen and Cecilie Bay-Richter for their advice, and precious comments and discussions on the topics of this paper. Also, we thank Alexandra Vossen for inspiring debates on statistics and gender neutrality of unicorn names alike.

Authors contributions. SA and AE conceived the idea, drafted and revised the manuscript and designed all figures. GW and AML provided critical feedback and helped to shape the revised version. All authors reviewed the manuscript for the intellectual content and approved the final version for submission.

Financial support. The authors received no specific funding for this paper.

Conflict of interest. Gregers Wegener is the Editor-in-Chief of *Acta Neuropsychiatrica* at the time of submission, but was not involved and actively withdrew during the review and decision process of this manuscript.

References

Aarts AA, *et al.* (2015) Estimating the reproducibility of psychological science. *Science* **349**(6251), aac4716. doi: 10.1126/science.aac4716.

Belzung C (2014) Innovative drugs to treat depression: did animal models fail to be predictive or did clinical trials fail to detect effects. *Neuropsychopharmacology* **39**(5), 1041–1051. doi: 10.1038/npp.2013.342.

Bert B, Heinl C, Chmielewska J, Schwarz F, Grune B, Hensel A, Greiner M, Schönfelder G (2019) Refining animal research: the animal study registry. *PLoS Biology* **17**(10), e3000463. doi: 10.1371/journal.pbio.3000463.

Biesecker LG (2013) Hypothesis-generating research and predictive medicine. *Genome Research* **23**, 1051–1053. doi: 10.1101/gr.157826.113.

Blokland A, *et al.* (2012) The use of a test battery assessing affective behavior in rats: order effects. *Behavioural Brain Research* **228**(1), 16–21. doi: 10.1016/j.bbr.2011.11.042.

Bonapersona V, *et al.* (2021) Increasing the statistical power of animal experiments with historical control data. *Nature Neuroscience* **24**(4), 470–477. doi: 10.1038/s41593-020-00792-3.

Brian Haynes R (2006) Forming research questions. *Journal of Clinical Epidemiology* **59**(9), 881–886. doi: 10.1016/j.jclinepi.2006.06.006.

Button KS, *et al.* (2013) Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* **14**(5), 365–376. doi: 10.1038/nrn3475.

Christley RM (2010) Power and error: increased risk of false positive results in underpowered studies. *Open Epidemiology Journal* **3**, 16–19. doi: 10.2174/1874297101003010016.

Crawley JN (2007) What's Wrong With My Mouse? Behavioral Phenotyping of Transgenic and Knockout Mice, 2nd edn. Hoboken, NJ: John Wiley & Sons, Inc. 10.1002/0470119055.

Dienes Z (2014) Using Bayes to get the most out of non-significant results. *Frontiers in Psychology* **5**, 781. doi: 10.3389/fpsyg.2014.00781.

Ellenbroek B and Youn J (2016) Rodent models in neuroscience research: is it a rat race? *DMM Disease Models and Mechanisms* **9**(10), 1079–1087. doi: 10.1242/dmm.026120.

Festing MFW (2020) The, completely randomised, and the, randomised block, are the only experimental designs suitable for widespread use in pre-clinical research. *Scientific Reports* **10**(1), 1–5. doi: 10.1038/s41598-020-74538-3.

Garner JP (2014) The significance of meaning: why do over 90% of behavioral neuroscience results fail to translate to humans, and what can we do to fix it? *ILAR Journal* **55**(3), 438–456. doi: 10.1093/ILAR/ILU047.

Giner-Sorella R (2018) Powering Your Interaction, Approaching Significance [Blog]. Available at: https://approachingblog.wordpress.com/2018/01/24/powering-your-interaction-2/ (accessed 12 November 2021).

Homberg JR, *et al.* (2021) The continued need for animals to advance brain research. *Neuron* **109**(15), 2374–2379. doi: 10.1016/J.NEURON.2021.07.015.

Ioannidis JPA (2005) Why most published research findings are false. *PLoS Medicine* **2**(8), e124. doi: 10.1371/journal.pmed.0020124.

Jennions MD and Møller AP (2003) A survey of the statistical power of research in behavioral ecology and animal behavior. *Behavioral Ecology* **14**(3), 438–445. doi: 10.1093/beheco/14.3.438.

Karp NA, *et al.* (2020) A multi-batch design to deliver robust estimates of efficacy and reduce animal use – a syngeneic tumour case study. *Scientific Reports* **10**(1), 1–10. doi: 10.1038/s41598-020-62509-7.

Kerr NL (1998) HARKing: hypothesizing after the results are known. *Personality and Social Psychology Review* **2**(3), 196–217. doi: 10.1207/s15327957pspr0203_4.

Lakens D (2013) Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology* **4**, 863. doi: 10.3389/fpsyg.2013.00863.

Lazic SE (2018) Four simple ways to increase power without increasing the sample size. *Laboratory Animals* **52**(6), 621–629. doi: 10.1177/0023677218767478.

Michel MC, Murphy TJ and Motulsky HJ (2020) New author guidelines for displaying data and reporting data analysis and statistical methods in experimental biology. *Molecular Pharmacology* **97**, 49–60. doi: 10.1124/mol.119.118927.

O'Collins VE, *et al.* (2006) 1,026 Experimental treatments in acute stroke. *Annals of Neurology* **59**(3), 467–477. doi: 10.1002/ana.20741.

Percie du Sert N, *et al.* (2020) The ARRIVE guidelines 2.0: Updated guidelines for reporting animal research. *PLOS Biology* **18**(7), e3000410. doi: 10.1371/journal.pbio.3000410.

Poindron P, Callizot N and Piguet P (2008) Theoretical considerations on animal models. In New Animal Models of Human Neurological Diseases, vol. **2**. Basel: Karger Publishers, pp. 1–10. 10.1159/000117719.

Pound P, Ebrahim S, Sandercock P, Bracken MB and Roberts I (2004) Where is the evidence that animal research benefits humans? *British Medical Journal* **328**(7438), 514–517. doi: 10.1136/bmj.328.7438.514.

du Sert NP, *et al.* (2020) The arrive guidelines 2.0: updated guidelines for reporting animal research. *PLoS Biology* **18**(7), e3000410. doi: 10.1371/journal.pbio.3000410.

Smith AJ, Clutton R E, Lilley E, Hansen KEA and Brattelid T (2018) PREPARE: guidelines for planning animal research and testing. *Laboratory Animals* **52**(2), 135–141. doi: 10.1177/0023677217724823.

The Roslin Institute - Training (2016) [Youtube Playlist], Introduction to Experimental Design. Available at: =https://www.youtube.com/playlist?list=PLbyRmcun-gis3nonV8IeVFRYjlpvpCufm (accessed 26 May 2021).