Reproducibility of animal research in light of biological variation

Bernhard Voelkl[1], Naomi S. Altman[2], Anders Forsman[3], Wolfgang Forstmeier[4], Jessica Gurevitch[5], Ivana Jaric[1] , Natasha A. Karp[6], Martien J. Kas[7],Holger Schielzeth[8], Tom Van de Casteele[9] and Hanno Würbel[1]

Nature reviews. Neuroscience, 21(7), pp. 384-393. Springer Nature 10.1038/s41583-020-0313-3

[1]Animal Welfare Division, Vetsuisse, University of Bern, Bern, Switzerland
[2]Department of Statistics, The Pennsylvania State University, University Park, PA, USA
[3]Department of Biology and Environmental Science, Linnaeus University, Kalmar, Sweden
[4]Department of Behavioural Ecology and Evolutionary Genetics, Max Planck Institute for Ornithology, Seewiesen, Germany
[5]Department of Ecology and Evolution, Stony Brook University, Stony Brook, NY, USA
[6]Data Sciences & Quantitative Biology, Discovery Sciences, R&D, AstraZeneca, Cambridge, UK
[7]Groningen Institute for Evolutionary Life Sciences, University of Groningen, Groningen, the Netherlands
[8]Institute of Ecology and Evolution, Friedrich Schiller University Jena, Jena, Germany
[9]Statistics and Decision Sciences, Janssen R&D, Johnson & Johnson, Beerse, Belgium

e-mail: hanno.wuerbel@vetsuisse.unibe.ch

Abstract | Context-dependent biological variation presents a unique challenge to the reproducibility of results in experimental animal research, because organisms' responses to experimental treatments can vary with both genotype and environmental conditions. In March 2019 experts in animal biology, experimental design and statistics convened in Blonay, Switzerland to discuss strategies addressing this challenge. In contrast to the current gold standard of rigorous standardisation in experimental animal research, we recommend the use of systematic heterogenisation of study samples and conditions by actively incorporating biological variation into study design, through diversifying study samples and conditions. Here, we provide the scientific rationale for this approach in the hope that researchers, regulators, funders and editors can embrace this paradigm shift. We also present a roadmap towards better practices in view of improving the reproducibility of animal research.

[H1] Introduction

Since the 17th century, the ability to reproduce research findings has been the acid test by which scientists distinguish facts from mere anecdotes[1]. Reproducibility — here defined as the ability to produce similar results by independent replicate studies — is thus a cornerstone of scientific methodology. Recent investigations, however, have shown that the reproducibility of research findings is poor across virtually all disciplines of research[2–9]. It is crucial to identify the causes of poor reproducibility and implement effective strategies for improvement for scientific, economic and ethical reasons.

The reproducibility of preclinical research involving animal models is deemed to be especially poor[10]. More than half of the published findings in this area are considered irreproducible, representing a cost of US$28 billion per year in the United States alone[11]. Additional resources are used on often fruitless follow-up studies, which in turn generate opportunity costs by preventing researchers from following more promising research avenues, or leading to time lost in scientific dead ends. These economic and scientific costs are associated with significant ethical issues. In biomedical research, poor reproducibility not only attenuates medical progress but also harms animals subjected to inconclusive studies and potentially puts patients who are enrolled in clinical trials at risk.

Current discussions about the causes of poor reproducibility in animal research mainly focus on violations of good research practice, including a lack of scientific rigor, low statistical power, analytical flexibility (for example, p-hacking) and publication bias[2,5,12]. In this Perspective article, we argue that, beside violations of good research practice, a major cause of poor reproducibility in animal research is a persistent disregard for the nature of biological variation in study design. Below, we explain where biological variation comes from, how it differs from random noise and why it causes issues with reproducibility. We then discuss why current research practice is inadequate for dealing with biological variation and call for a paradigm shift in experimental design to improve reproducibility in animal research. Specifically, we propose diversification of study subjects through deliberate heterogenisation of environmental factors as a measure of good experimental design.

[H1] Biological variation

[H2] ***Sources of biological variation.*** Variation is ubiquitous in nature and even casual observations reveal that individual organisms differ in numerous phenotypic traits. Such phenotypic variation reflects the combined effects of the organisms' genotypes and their responses to the environment, integrated over their lifetimes[13–16]. Phenotypic variation covers all levels of organization from molecular

mechanisms to behaviour. There are many biological causes of phenotypic variation besides genetic differences, including developmental stage or age, early experience and social status. Variation owing to the environment is complex and varies with time, spatial scale (for example, climate) and the nature of environmental factors (for example, food, predators, mates and environmental toxins). The norm of reaction describes the relationship between one or more environmental factors and the phenotype for a given genotype[17,18], and such norms of reaction may differ among genotypes (BOX 1). Thus, it is not uncommon for different genotypes to respond differently to environmental factors[19]. The effects of environmental factors accumulated over a lifetime may not be easily detectable by morphological, physiological or behavioural analysis, although they may leave a unique fingerprint on gene and protein expression levels, thereby contributing to the fine-tuning of the phenotype[20]. Recent advances in the study of epigenetics have added a layer of complexity to our understanding of the interactions between genotype and environment in the expression of phenotypic plasticity[21,22].

In experimental animal research, the effect of a treatment is typically measured at the level of the phenotype. It can be thought of as the plastic response of some phenotype (the animal model) to the experimental treatment. Without plasticity in response to a treatment (for example, the administration of a drug or a genetic manipulation), there would be no treatment effect (that is, effect size = 0). However, the direction and magnitude of a treatment effect depends not only on the nature, duration and intensity of the treatment, but also on the animal's current phenotype and the experimental context[23,24]. As phenotypes are complex and influenced by many interacting factors, the effect of the independent variable (for example, the experimental treatment) on an outcome variable (the dependent variable) is also context-dependent. Thus, experimental results vary with both the internal state of animals (determined by genotype and experiences throughout development) and the external environmental factors (the environment in which the experiment is conducted). Environmental factors may interact additively or synergistically with the internal state of the animals, shaping their responses to the experimental treatment in specific ways (FIG. 1).

[H2] ***Biological variation and current best practice.*** In laboratory animal research, current best practice for dealing with biological variation is strict standardization of both the animals and their environment[25–27]. Standardization in animal experimentation has been described as "the defining of the properties of any given animal (or animal population) and its environment, together with the subsequent task of keeping the properties constant"[25]. It is intended, first, to reduce within-experiment variability in order to increase statistical power and, second, to reduce between-experiment variability in order to "increase

the reproducibility of group mean results from one experiment to another", thereby "improv[ing] comparability of results within and between laboratories"[25]. "The defining of the properties" does not necessarily implicate identical environmental conditions for all animals of a study population, and other definitions of standardization exist that refer to 'the setting of, and compliance with, standards' rather than making everything the same (for example, see REF.[28]). However, in laboratory animal experimentation, standardization is generally equated with such homogenization[29–32], and throughout this article, the term standardization refers to the homogenization of study populations. Standardization renders animals within experiments more homogenous and thus less variable. Reduced variation in the results increases statistical power and allows a reduction of sample size (to detect a given effect size). Therefore, standardization has been advocated also for ethical reasons as a means of reducing animal use as required by the 3Rs principle (Replace, Reduce, Refine)[31,33–36].

There are two main problems with this conception of standardization as applied to laboratory animal research. It is based on the confusion of biological variation with extraneous noise and on the myth of a pure treatment effect that 'emerges' as more sources of variation are eliminated. Whereas standardization can be an effective means to reduce extraneous noise (for example, measurement error and undesirable environmental effects), it fails to address biological variation. Since variation is a fundamental property of any population of organisms, treatment effects can only be assessed and interpreted meaningfully against biological variation — including gene × environment interactions. Owing to context-dependent variability in responses to treatment (BOX 1), there is no such thing as a pure treatment effect for a population of living organisms. Any definition of a target population, therefore, needs to consider the range of genotypic and environmental variation for which the inferences of a study should be valid (the inference space). Studies that are too narrowly defined cannot reliably be generalized: if only males are included, the results may differ in meaningful ways in females[37,38]; the responses of a single inbred strain may not hold for other strains[39,40]; and mice housed in isolation might respond differently to certain drug treatments than individuals housed in groups[41]. Although extension of the inference space has been discussed specifically with regard to genetic variation and the inclusion of both sexes in preclinical animal studies (BOX 2 and BOX 3), here we argue that this discussion should be extended to diversification of environmental conditions.

Outcomes, both the main effects of treatments and treatment × environment interactions, that are stable under large biological variation are considered to be robust[42], and may be characterized by the same flat norm of reaction for all genotypes and all variants of environmental factors (BOX 1 and FIG. 1). However, such cases of universal robustness are probably rare exceptions rather than the rule.

In most cases, treatment effects will vary depending on a set of both genetic and environmental factors. Such modulating effects can be highly specific and unexpected. For example, a change from open cages to individually ventilated cages (IVCs) altered outcomes in a mouse model of infection-mediated neurodevelopmental disorders[43], the behavioural sensitivity of wild-type mice[44,45] , and the behavioural phenotype of a validated mutant neuregulin 1 mouse model for the schizophrenia[46] , but not the behavioural phenotypes of three commonly used inbred mouse strains[47] . Knowledge about context-dependent variation of treatment effects is a crucial aspect of scientific evidence. It is necessary for identifying the target population, as well as the conditions under which a finding is likely to be reproducible[24]. It is also key for translational research and the very basis of precision medicine[23,48–50].

[H2] ***Reproducibility and the standardization fallacy*****.** Reproducibility is assessed by comparing the results of independent replicate studies[12,51]. The conditions of any two studies are never exactly the same, even when researchers go to great lengths to harmonize characteristics of animals, housing conditions, experimental protocols and test conditions[24,51–53]. Differences are unavoidable since the animals, the personnel interacting with the animals, the animals' microbiome and many other factors resist standardization[39,54–60]. Different laboratories, therefore, inevitably standardise the variables to different local study contexts, producing increasingly distinct study populations as standardization gets more rigorous. With every additional variable that is standardized, one risks that the inference space of a study (and with it the external validity of its results) decreases[29,61,62]. This misguided attempt to enhance reproducibility at the expense of external validity is referred to as the standardization fallacy[63]. Although direct evidence for the standardization fallacy is currently limited to simulations across replicate studies[29,62,64] and only a few dedicated experimental studies[65,66], there is indirect evidence showing, for example, that the experimenter or the laboratory may account for most of the variation in outcome measures across replicate studies within or between laboratories, respectively[23,52].

Results can only be reproduced successfully if they are robust against the variation that exists between independent replicate studies. It is therefore not surprising that standardization has invariably been found to be a cause of, rather than a cure for, poor reproducibility[65–68] (but see REF.[69], and REF.[70] for a critical analysis of REF.[66]). Eliminating biological variation through the use of standardization to narrow the inference space of a specific animal phenotype may, therefore, be a highly inefficient strategy for generating scientific evidence. It is akin to the atomization of animal research by investigating each specific genotype × environment interaction in a separate experiment, thereby minimizing the information gain per experiment to virtually zero. The detection of robust and

reproducible effects of interventions would thus require a very large number of independent replicate studies and rely entirely on meta-analysis. The other extreme, however, is not an efficient strategy either. Incorporating the full range of both genetic and environmental variation into the design of every experiment would render experiments unmanageable. A key challenge for future research is thus to find the right balance between biological complexity and experimental tractability. The following section presents approaches to account for biological variation in view of the limitations set by these two extremes.

[H1] Call for a paradigm shift

In contrast to the current practice of dogmatic standardization, we advocate systematic heterogenization of animal subjects by deliberately incorporating known sources of biological variation in study designs. Heterogenization may be based on controlled variation, for instance by systematically varying genotype (for example, both sexes or several inbred strains), state and history of the individual (for example, different housing conditions or different age classes, or test condition (for example, different test times or alternative test systems). Alternatively, heterogenization may be based on uncontrolled variation, for example, by using outbred study populations, by splitting experiments into multiple independent batches of animals, or by conducting multi-laboratory studies. These different types of heterogenization, as well as rigorous standardization, have their place in research, as outlined in more detail below.

[H2] ***Study designs and analysis plans.*** Study design is often taught as if each experiment was a fully independent and conclusive study. However, most experiments are part of research programmes including a series of experiments, each providing incremental gains of knowledge that guide the next steps in the programme[71,72]. Ideally, research into new and unexplored areas begins with exploratory studies that can be used to generate and select hypotheses worthy of further investigation[73,74]. Such hypotheses may then be tested in confirmatory studies to establish proof of concept, followed by studies assessing the generalizability of the findings. However, often there is no clear distinction between exploratory and confirmatory studies[49]. This can cause problems as different types of study and different stages of research require different study designs, sample sizes, analysis plans and interpretation of outcomes.

Initial exploratory studies are usually small, limited to a single strain of animals and often only one sex (predominantly males in animal experiments[75–77]), and they are frequently conducted under

rigorously standardized conditions. Given their aim to generate new hypotheses or identify hypotheses worthy of further investigation, this is a highly inefficient strategy, more likely to generate 'findings' that are context specific. Exploratory studies based on carefully heterogenized designs, however, may provide considerable knowledge about how the effect of the experimental intervention (the effect size) is modified by the heterogeneous features (including both genetic and environmental factors) being incorporated in the experimental design. Knowing whether effect sizes are likely to be robust or context-dependent permits a much more targeted approach to follow-up studies testing for proof of concept and generalizability[49].

There are various ways to heterogenize a study population. For example, we may want to estimate an average effect without exploring the impact of each heterogenization factor (for example, strain or environmental condition). In this case, we may split the study sample into groups or 'blocks', using a randomized complete block (RCB)[78] design (BOX 4). This usually does not require increasing the sample size compared with a completely standardized study design to achieve the same power[79]. In many cases, there is already a blocking factor present in the study design, for example, to account for batch, cage or pen effects. In such cases, heterogenization may be achieved by deliberately adding additional heterogeneity between blocks, which improves external validity without sacrificing the internal validity achieved by within-block standardization. Such block heterogenization is suitable to determine whether a treatment effect is robust over a range of conditions, in which case it is also more likely to be reproducible across studies than an effect that interacts strongly with a blocking factor.

Sometimes we are interested in identifying the sources of biological variation modulating the response to the treatment and assessing the magnitude of the influences of specific factors (for example, sex, age or specific environmental parameters), rather than just maximizing external validity. In such cases, these factors need to be included as fixed effects (differences in the means owing to the influence of independent variables) in the experimental design and analysis. The inclusion of fixed effects as factors in the study design, especially if they are varied across multiple factor levels (that is, values), may require larger sample sizes than standardization or heterogenization using a random blocking factor (a factor increasing variability). Therefore, this should be considered for cases only where the estimation of these effects is scientifically warranted, for example to assess the effects of sex — which we generally recommend — or other relevant biological variables (for example, specific comorbidities in animal models of diseases) on the outcome variable.

Heterogenized study designs, which incorporate biological variables either as random or fixed effects, should become the default option for almost all study types of experimental animal research —

including exploratory studies. Rigorous standardization of study animals to a single genotype and a single sex and to being kept under one specific environmental condition can only be justified on the grounds that either the outcome of interest was previously shown to be robust against variation in these factors (albeit absence of evidence should not be mistaken for evidence of absence) or the research question is truly limited to that specific context (for example, as in the study of sex-specific diseases). In all other cases, systematic heterogenization will be scientifically more valid and — especially when considering single studies as parts of a larger research programme — will also be economically beneficial and ethically preferable.

[H2] ***Scientific, economic, and ethical implications*****.** To assess the scientific, economic and ethical implications of heterogenization, it is important to take a perspective that extends beyond the individual experiment[80]. By reducing within-experiment variation, standardization can increase test sensitivity for a specific standardized study context, which in turn allows reducing sample size as required by the 3Rs principle[34]. However, as standardization produces results with a validity that may be limited to that specific context, it generates a greater need for follow-up studies, thus requiring the use of additional animals. If we seek to minimize the use of animals to achieve our research goals, we should focus on maximizing the amount of knowledge gain per animal and/or per study rather than minimizing the number of animals per study. The scenarios presented above demonstrate how scientific evidence can be generated more efficiently and, as a consequence, more ethically, if biological variation is accounted for in the study design from very early on[81,82]. It is time to update the textbooks of laboratory animal science and establish systematic heterogenization of study populations as a new standard. As this implies a true paradigm shift, change management towards better practice is needed.

[H2] ***The path to implementation*****.** Gene × environment interactions, phenotypic plasticity and reaction norms are fundamental biological concepts and standard knowledge taught in undergraduate genetics and biology classes (BOX 1). The same applies to block experimental designs to incorporate and control for heterogeneity when studying the effect of one or more independent factors upon an outcome variable (BOX 4). Moreover, the limitations of standardization for external validity and reproducibility of results from animal experiments have long been known[53,61,63,83]. Why then do laboratory animal scientists persist in promoting a principle of experimental design — rigorous standardization — that is incompatible with these insights? Answering this question requires a consideration of the forces

encouraging change and the resisting forces that hinder researchers from embracing biological variation as part of their experimental paradigms and thus maintain the status quo (standardization).

Table 1 lists forces impacting researchers' engagement with changing practice. Understanding these forces and considering the interplay highlights that resisting forces dominate, which explains the challenge for our community to achieve the paradigm shift that is needed. A closer look at these factors will highlight how we can unfreeze the status quo by strengthening the driving forces and weakening the resisting forces, allowing the paradigm shift to occur[84].

[H2] ***Exploring resisting forces and potential solutions.*** Our current research culture is made up of the beliefs, values and norms of behaviour (protocols and systems) of the community of researchers. A central belief determining current research practice is the conviction that standardization is a universal means to improve the validity of animal experiments and meet our ethical obligations to use as few animals as possible. As we have outlined above, this assertion does not hold if standardization is removing relevant biological variation. Although this problem has been identified previously, standardization is culturally embedded in our community as the norm and best practice. It is done without questioning its validity. Consequently, more advocacy will be needed to convey the Janus-faced nature of standardization to the wider research community. One roadblock is conflicting evidence from other scientific disciplines like physics, in which standardization is indeed an effective measure to reduce measurement error of technical replicates and, hence, to improve both the internal and external validity of study results. However, heterogenization is commonly accepted practice in many other biological disciplines, particularly those dealing with whole organisms, including quantitative genetics, animal and plant breeding, ecology and evolutionary biology. In order to overcome this resistance, we need to challenge the underlying beliefs that standardization is best practise and to promote awareness that biological variation of the phenotype differs fundamentally from random noise (as exemplified in BOX 1 on reaction norms).

The designs recommended here introduce challenges through changes in the way we practically run the experiments and analyse the data. A significant blocker for our community to embrace such changes is the current norm to publish studies with a narrow inference space with no acknowledgement of the limitations of the study findings. This approach has a significant impact on our research culture, as we are rewarded for publications regardless of the robustness of our findings. A further obstacle to change is the argument that heterogenization increases the complexity of the experiment and, thereby,

complicates the analysis and increases the required sample size and economic costs of the experiment. Although it is true that heterogenized study designs are more complex, this does not hinder analysis, because statistical tools to deal with the added complexity are readily available (see BOX 3 on blocking). Here, scientists might need more encouragement to engage with those statistical concepts and apply them in their research practices[85].

Despite a general understanding of the problem, researchers who wish to implement heterogenization face several unknowns. Which factors will have the strongest effects on the overall variation? How strong will the effects be? Within which range should we vary environmental factors? How different should the experiments be to be considered independent replicates? When does this require replication in a different laboratory, and when are sequential batches within the same laboratory sufficient? For some well-researched treatments or compounds we might have information about the effect of some of the more common heterogenization factors like sex or strain, but in most cases we will lack this information and cannot answer those questions upfront. Some answers can be gleaned from the literature on the evolutionary biology and ecology of related animals, which provides a rich source of information on phenotypic and genetic heterogeneity. Educated guesses and rules of thumb for certain groups of interventions might give some guidance, but in the end, we have to accept intrinsic uncertainties that can only be resolved empirically. Further research is therefore essential to explore these issues and provide guidance to the community in practical steps that can be taken.

Given the uncertainties with regard to heterogenization factors that will prove effective for any specific treatment, we do not think that the way forward should be a list of compulsory factors to be heterogenized in every study. Instead, we recommend that heterogenization of sex, genotype, age and environmental conditions should be recommended in general terms and that experimenters should be asked to discuss their choice of heterogenization measures — or the lack of such measures — with respect to the intended inference space.

[H2] ***Overcoming the reproducibility crisis.*** Besides counteracting factors inhibiting change, there is also a need to strengthen those factors motivating and driving change. Arguably the most compelling one is the reproducibility crisis in biomedical research. Reports pointing out issues with reproducibility have accumulated over the years and the desire to solve this crisis should be a very strong motivation for driving change. Improving reproducibility can reduce long-term research costs, increase efficiency of drug development and reduce suffering of animals used in inconclusive research.

Connected to poor reproducibility, there is a related and very compelling issue: it is the reputation of research itself — and of animal research in particular — that is at stake[4,86]. The public funds research on the principal understanding that researchers use funding resources judiciously and efficiently. If the research community fails to resolve the current reproducibility crisis, then the public might legitimately question whether researchers adhere to this societal contract and whether investment in this kind of research should continue. Along the same lines, the right to use animals for research that might inflict pain and suffering to the animals is a privilege granted to researchers by society on the understanding that their research benefits humanity and that researchers use the animal resources responsibly, avoiding unnecessary harm and suffering. Again, a failure of the scientific community to resolve the current crisis might instigate a discussion whether this privilege should be revoked. Here, we believe that it is important to communicate that ignoring or denying the existence of a major reproducibility problem is not a solution and that only an honest and serious attempt by the entire scientific community to solve this problem can secure a continued trust in science by the general public.

With respect to sample size, recent studies indicated that heterogenization can be introduced without a need to increase the overall sample size[53,67,87]. Larger sample sizes are only needed when multiple factors are heterogenized; however, the increase in economic and ethical costs of larger experiments should be more than outweighed in the long run, as fewer follow-up studies will be required and fewer dead ends will be pursued. Promises for long-term benefits are, by their very nature, vague, which means that they are rather weak arguments for implementing change. However, if the focus is shifted from the costs of an individual study (both financial and ethical in terms of numbers of animals used) to the amount of knowledge gained per study or per animal (Table 1), then it becomes immediately apparent that heterogenized studies can deliver a better benefit/cost ratio than narrowly standardized studies. This change of focus from the number of animals within single experiments to the value delivered by these experiments is also reflected in the recent change in the definition of Reduction by the National Centre for the Replacement, Refinement and Reduction of animals in Research (NC3Rs), which now includes “experiments that are robust, reproducible and truly add to the knowledge base”[88]. Researchers, regulators, funders and editors need to understand that studies allowing inferences about both sexes, genetically diverse populations or a variety of environmental conditions will add more richly to the knowledge base, and that standardization of animal subjects inevitably reduces the inference space. A promising way forward might be insistence by editors, reviewers and funders that authors have to specify the inference space of their studies; that is, the population — and the biological variation

within that population — about which they will be able to draw inferences. Some funders are already requiring some forms of heterogenization (BOX 2 and BOX 3). For example, both the US National Institutes of Health (NIH)[89] and the Organisation for Economic Co-operation and Development (OECD)[90] recommend that studies in preclinical biomedical research should comprise both sexes and the European Medicines Agency (EMA) requires that animal toxicity tests for compounds have to be made in at least two different species, prior to translation to human subjects[91]. We therefore propose that reporting guidelines (for example, the ARRIVE guidelines[92] and Nature Research's Reporting Checklist For Life Sciences Articles[93]) request that experimenters explicitly state the intended inference space and discuss their results with respect to the measures taken (for example, the factors that were heterogenized) to cover that inference space.

[H1] Conclusions

Accumulating evidence of poor reproducibility of research has stimulated heated debate about possible causes and remedies of irreproducibility leading to the so-called reproducibility crisis. Suggested causes of poor reproducibility include lack of scientific rigor, low statistical power, publication bias, analytical flexibility (for example, p-hacking), pseudoreplication and outright fraud[6,12,94–96]. These causes are all thought to be promoted by a system that rewards novel and spectacular findings — even if they are spurious — more than robust, reproducible evidence[97]. However, we contend that this list is incomplete for research involving living organisms as an important source of replication failure has been neglected: standardization of the animals, leading to unrealistically low estimates of biological variation and, as a consequence, to study-specific, idiosyncratic results. As biological variation differs from random noise and variation of technical replicates — as clearly demonstrated by the reaction norm framework — its removal through standardization generates overconfident and biased estimates. Here, we have outlined the rationale for our claim, as well as its scientific, ethical and economic implications, and presented targeted scenarios for improvement. We maintain that unless researchers take the context-dependency of their animals' treatment responses into account, reproducibility of animal research will remain limited despite efforts to avoid other causes that affect reproducibility. We call on the community (researchers, publishers, policy makers, professional bodies, funders etc.) to engage and explore how they can support the paradigm shift that is needed to deliver robust research.

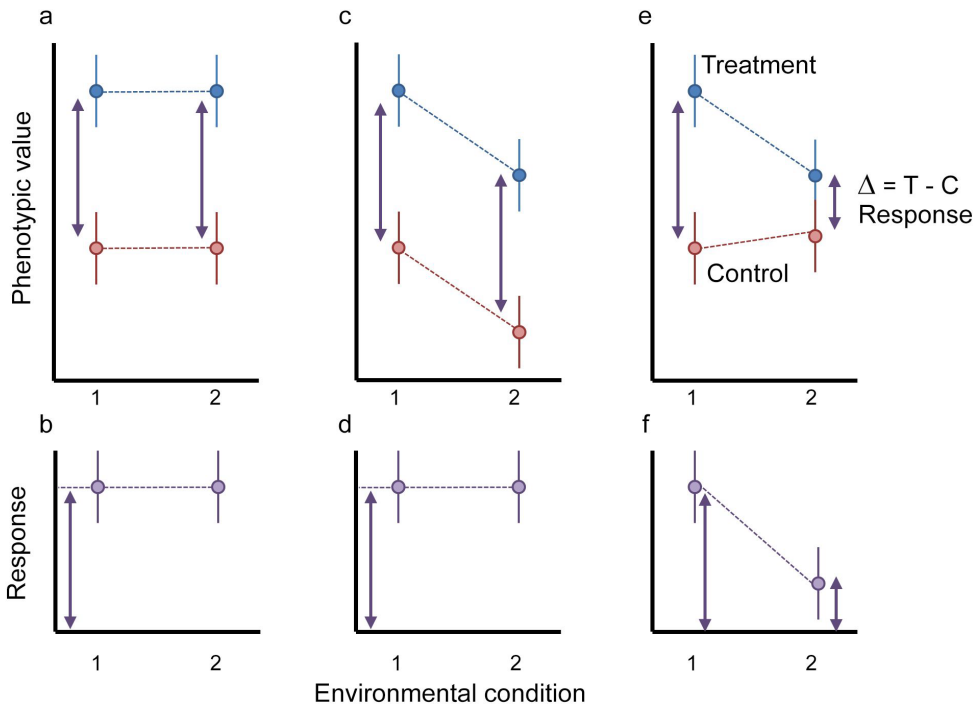
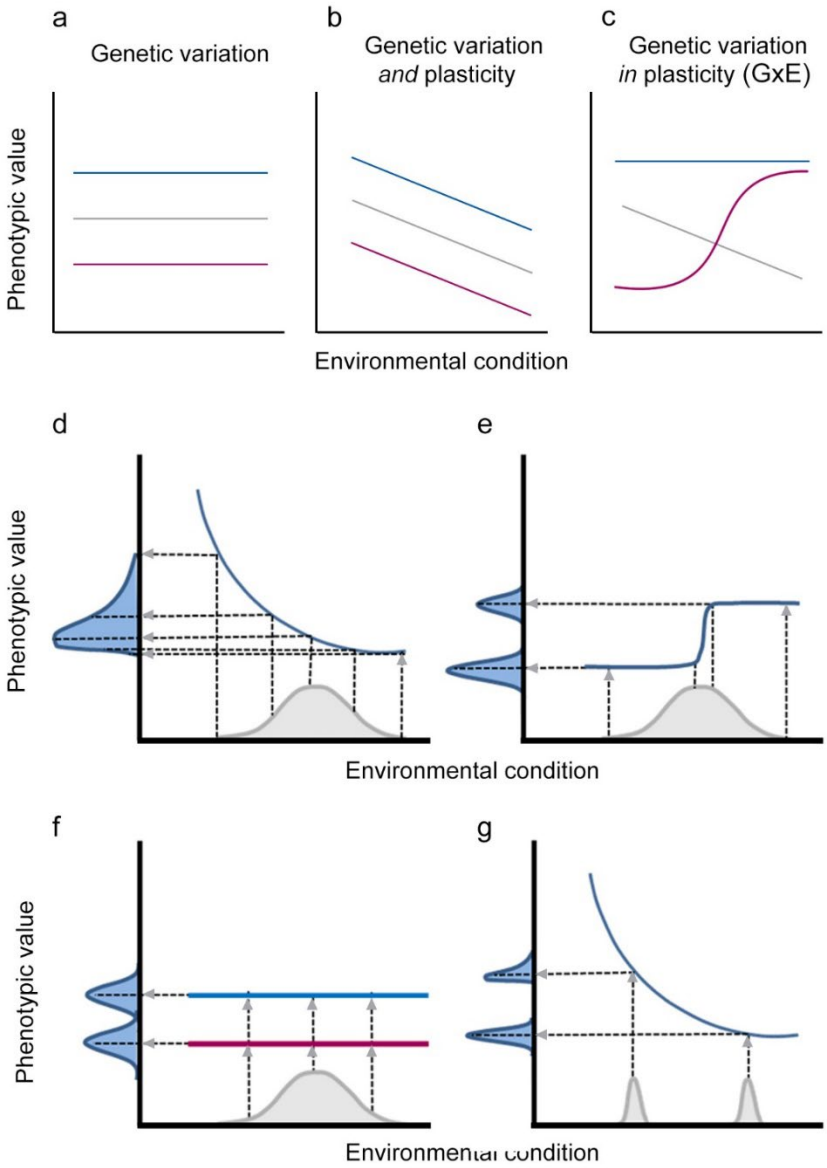


Figure 1 | **Context-dependent treatment effect**. **a** | Phenotypic values measured in treatment and control animals can be robust, which means they are insensitive to changes in the environmental condition. **b** | In this case, the response to the treatment — that is, the treatment effect (double-headed arrows) — is also robust. **c,d** | If phenotypic values are sensitive to environmental conditions, the response can still be robust if the environmental effect is purely additive. **e,f** | However, a response to a treatment can be context dependent if there is an interaction between the treatment and the environment.

Box 1 | **Genetic variation, plasticity and norms of reaction**

Phenotypic variation among and within individuals typically reflects the combined effects of genetic differences and environmentally induced variation[13–16]. When there is no plasticity (that is, the norms of reaction are flat), phenotypic differences among genotypes are robust across environmental conditions (see the figure, part **a**; the blue, grey and red lines in parts **a**–**c** represent distinct genotypes). However, the relative importance of environmentally induced phenotypic variation is highly context dependent and typically varies among populations, traits, and genotypes. When the norms of reaction for different genotypes have parallel positive or negative slopes (see the figure, part **b**), the plastic responses induced by the environment are shared and phenotypic distributions reflect the combined effects of genetic and environmental variation. The sensitivity or responsiveness of phenotypic expression can vary among environmental components. A given phenotypic trait may show a plastic response to some environmental factors while being insensitive to others. Similarly, a given environmental factor may induce a plastic response in some phenotypic trait or traits while the development of other traits may be robust and independent of the same factor. The phenotypic response induced by an environmental factor can also be genotype specific, in which case the phenotypic variation in a population depends on the joint effects of genetic variation, phenotypic plasticity, and genetic variation in plasticity. When the reaction norms vary among genotypes there is genetic variation in plasticity (that is, genotype × environment interactions (G×E)), meaning that the plastic response induced by the environmental factor varies according to genotype (see the figure, part **c**).



The norm of reaction describes, for a specific genotype, how the distribution of an environmental factor is translated into a phenotypic distribution. This means that even in genetically homogeneous populations, such as inbred laboratory strains, the patterns of phenotypic variation can vary depending on the environment. A continuous, normally distributed environmental variation can generate, for example, a continuous, skewed phenotype distribution (see the figure, part **d**). However, for threshold traits with a step-shaped reaction norm, continuous environmental distributions can also generate discrete (categorical) or bimodal phenotypic trait distributions for a single genotype (see the figure, part **e**). Here, the expression of a phenotypic trait changes from one state to another at some critical level, dosage, intensity or concentration in the environment. The critical level that induces the phenotypic shift from one state to another (for example, response or no response) may vary among

genotypes. In other cases, bimodal or multimodal trait distributions may manifest in populations that comprise different genotypes, regardless of whether they show or do not show plasticity (see the figure, part **f**). Furthermore, a given phenotypic trait may display a discrete or bimodal frequency distribution if the genotype or genetically homogeneous strain is exposed to a discrete or bimodally distributed environment (see the figure, part **g**). A practical implication of such context-dependent responsiveness is that the phenotypic responses induced by a specific experimental treatment (for example, intervention studies designed to evaluate drug responsiveness) may vary between trials conducted in different laboratories. The importance and consequences of developmental plasticity, phenotypic flexibility and genotype by environment interactions are well established in quantitative genetics and evolutionary ecology, and can explain why different studies may generate conflicting outcomes.

Box 2 | **Heterogenization in animal research**

Genetic heterogenization

Soon after the creation of inbred strains of rodents, researchers began to debate the advantages and disadvantages of their use as models for human medical conditions. Proponents for the use of inbred strains mainly emphasize the advantage of working with a genetically well-defined and standardized model[35,87,98]. A stringent breeding regime over 20 or more generations will lead to an inbreeding coefficient larger than 0.99 and homozygosity in over 98 percent of all loci[99], making animals of one strain from one breeding line almost genetically identical (though a few de-novo mutations, tandem repeats and transposon insertions always add marginal variability[16]). It has been noted that reliance on a single genotype can be risky as a sample of a single inbred strain will not reflect the genetic diversity of natural populations to which the insights should be applied in the end[100]. Furthermore, homozygosity as a result of inbreeding might render inbred mice poor models for outbred populations of heterozygous organisms. Five different approaches for genetic diversification within an experiment have been suggested: use of outbred strains[100], F1 hybrids[101], diversity outbred strains[102], multiple inbred strains[35] and both sexes[79,103,104]. The choice of the heterogenization strategy will depend on whether one aims exclusively for variation within individuals (that is, re-establishing heterozygosity through hybridization), variation between subgroups of individuals (use of both sexes or multiple strains), or variation between individuals (use of outbred strains).

Given the genetic uniformity of inbred strains, one might expect to find less between-animal variation of phenotypes in inbred strains than in stocks of outbred or wild-derived mice. The empirical evidence for this assertion is mixed and some empirical studies[105,106] and a recent meta-analysis of 241 data sets[107] report no overall difference in phenotype variability between inbred and outbred strains. Furthermore, groups of inbred mice kept under the same standardized conditions show sometimes surprisingly large phenotypic variation[40]. The reasons for high variability in inbred strains are poorly understood, although it was suggested that heterozygosity might have stabilizing effects, buffering the development and ensuring robust phenotypes. The loss of heterozygosity due to inbreeding might then disrupt these buffering mechanisms, leading to unstable phenotypes highly susceptible to fluctuations of the internal and external milieu[40,107–109].

Other targets for heterogenization

Age affects many physiological and behavioural processes[110–112] and has been suggested as a feasible factor for heterogenization[53,65,79,83]. In addition to age, reproductive experience has been shown to influence diverse physiological parameters and epigenetic marks[113–116]. Furthermore, seasonal changes, differences in the light regime and differences in the timing of experiments have been shown to affect study outcomes[52,68,117,118]. These environmental factors could be considered as further heterogenization factors. An experimental study showed that co-housing laboratory mice with feral and pet store mice profoundly affected the immune system of the animals, instigating memory T cell differentiation and leading to substantial differences in immune responses to bacterial infection[119]. Heterogenizing the microbial environment of laboratory mice was suggested as a tool for producing models with immune responses resembling those of adult humans more closely. Only a few studies have used different housing conditions for heterogenization, such as cage size or environmental enrichment[65,66], possibly because this is logistically more demanding. However, an earlier study found that memory deficits in mice deficient in hippocampal NMDA-type glutamate receptors were overcome by environmental enrichment, possibly as a result of enrichment-induced NMDA receptor-independent synaptogenesis[120]. In this case, systematic variation of environmental complexity facilitated the detection of a biologically relevant gene × environment interaction.

Box 3 | **Inclusion of both sexes**

About three decades ago, an imbalance in clinical research, with female subjects being underrepresented, led to a series of policy changes to encourage or enforce the inclusion of women in medical studies[121,122]. Although those initiatives were originally restricted to late-phase (phase III) clinical studies, more recently the US National Institutes of Health (NIH), the Canadian Institutes of Health Research and the European Commission extended their recommendation for the inclusion of both sexes to pre-clinical animal studies[77,123,124]. Sex-based differences in basic biological function, disease processes and treatment responses have been found in many animal models[79,125–132]. There are marked differences in global gene expression patterns between male and female animals. In mice, the majority (50–75%) of genes have been shown to be sex-biased (that is, expressed at different levels in the two sexes) even in non-reproductive tissues such as liver, fat, muscle and brain[133]. Cell-culture studies have demonstrated that neurons from male and female mice respond differently to various stimuli. Neurons from male mice were more sensitive to stress from reactive oxygen species and excitatory neurotransmitter, whereas neurons from female mice were more sensitive to some stimuli that prompt apoptosis[38,134]. These differences could have potential implications in treatments for stroke, traumatic brain injury, cerebral ischemia and other neurological or psychiatric conditions, such as Parkinson disease and schizophrenia. Apart from sex differences likely stemming from differences in X and Y chromosomal genes, sex-specific responses can also be mediated through hormones acting directly on genes throughout the genome[125,133]. As a consequence, researchers have started to diversify their study samples by including female animals, although parity has not been reached and specifically in neuroscience males are still predominant[75,76].

One of the most common concerns regarding inclusion of female animals in research is the fear that this will require larger sample sizes. This would increase not only the costs but also the workload for research and, consequently, slow down scientific progress[135–137]. Furthermore, owing to hormonal fluctuations across the reproductive cycle, female animals are believed to be more variable and therefore would inherently require larger sample sizes. However, recent meta-analyses that examined variability among male and female mice[41] and rats[138] showed that males were equally or even more variable in all measured parameters. Whether inclusion of both sexes requires a substantial increase in the sample size depends on the specific aim of the study. If separate subgroup analyses for the sexes are planned, a balanced factorial design can ensure that the sample size need not to be doubled but that a moderate increase of the required sample size will suffice[103,104]. Otherwise, if sex is added merely as a

heterogenization factor without the aim to test for sex-specific effects then this does not require larger sample size (or only a minimally larger sample size) than a single-sex experiment[79].

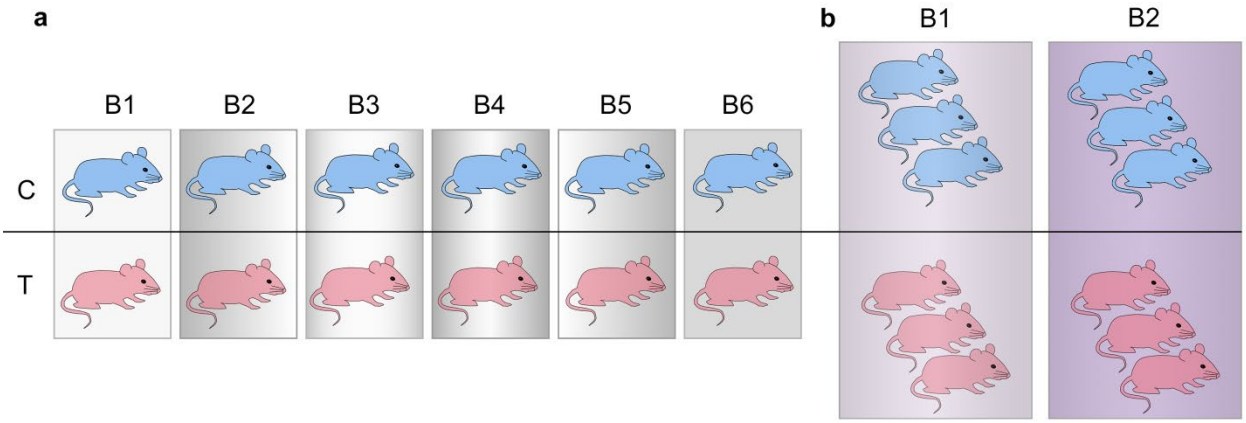
Box 4 | **Blocking design and heterogenization**

Blocking is an effective means of exploiting the benefits of both standardisation and heterogenization. Within blocks of subjects, the experimental conditions can be standardized as rigorously as possible (for example, use of same genotype, same age and same experimental context), so that any differences in response to the experimental treatments will most likely be attributable to the treatment. However, the blocks themselves can be heterogeneous and vary in one or several aspects. In the classic randomized complete block design (RCB) experiment, each treatment is assigned randomly to a single animal within each block (see the figure, part **a**). Such data can be analysed using a paired t-test between treatment and control that are paired within blocks (B1 to B6) when there are only 2 treatments (treatment (T) and control (C) in part **a** of the figure), or in a linear (mixed-effect) model where block is treated as a fixed or random effect when there are more treatments. The latter models have the advantage of being adaptable to more complex experimental designs, for example, blocks of time that are nested within blocks of laboratories.

The strength of the RCB design is that the treatment effect can be estimated within each block, and therefore it is independent of the block to block heterogeneity. Hence any context-general treatment effect will be unaffected by heterogeneity among blocks. Moreover, the same design can be used to explore context dependencies. We can include fixed effects that describe differences among blocks and their effects can be estimated. We can therefore determine treatment effects that are consistent across blocks (and hence are likely to be generalizable to more heterogeneous settings) as well as those that differ among blocks.

One problem with the classical RCB design is that treatments that have a consistent effect in only a subset of the blocks may not be identified if the estimated effects are highly variable between the remaining blocks. This can be mitigated in two ways. If there is prior knowledge on the sources of context dependencies, heterogeneity across candidate contexts can be built into the experiment. This can be analysed by fitting additional fixed effects for these factors using a split plot design where these factors are whole plot factors. If there is no prior knowledge about context dependencies, these can be explored and mitigated by replicating within blocks (for example, B1 and B2 in part **b** of the figure), thus yielding a measure of within block variability which can be used to assess whether treatment effects vary among blocks. Replication within block might be specifically of interest for late-phase studies, in which we do not only want to get a general proof of concept but also gain insights into the different sources of variation and their magnitude. It must be noted that replication within blocks is much less

effective in increasing power to detect context-independent effects (main effects) than use of additional blocks, but it is required to test for interactions between the factors of interest and the context factors.



Although the RCB is a highly efficient means of combining rigorous standardization with heterogenization, it requires that blocks are sufficiently large to include at least one replicate of each treatment. When this cannot be done for technical reasons or when a larger number of blocking factors are considered, incomplete block designs[139] are available that provide much of the same advantage with a small cost to the power of testing the treatment main effects.

Box 5 | **Forces driving and resisting change in experimental animal research**

Forces driving change

- Reproducibility crisis
- Ethical focus on knowledge gain per animal
- Long-term efficiency in resource and time use
- Reputation of in vivo research

Forces resisting change

- Scientific reward system favouring single small-scale studies
- Ethical focus on number of animals per study
- Belief in the value of standardization
- Cost per experiment
- Complexity of design and analysis
- Research culture (what is considered best practice)
- Unknown solution (how to profit from biological variation)

Glossary

Biological variation

Biological variation is the variation of phenotypes in a population of organisms. It is the result of genetic variation, environmental influences on the organism and gene × environment interactions.

Confirmatory studies

Confirmatory studies are designed to test specific hypotheses about the existence of a relationship or effect, its direction and magnitude, using inferential statistical methods. The hypotheses are based on previous knowledge of the study system.

Exploratory studies

Exploratory studies are designed to probe for relationships or treatment effects of novel interventions without specific hypotheses about the direction and size of effects. The outcome of an exploratory study is a descriptive account of the observed effects.

External validity

External validity is the extent to which findings can be generalized to the desired inference space of animals (including humans) and/or other environmental conditions.

Gene × environment interactions

These subsume the non-additive joint effect of genetic and environmental influences on the development of the phenotype. As a consequence, environmental influences can have different effects on the phenotype depending on the organism's genotype or genes can have differential effects depending on features of the environment.

Genotype

The genotype is an organism's hereditary information as encoded in the genome.

Heterogenization

Heterogenization is the deliberate augmentation of systematic or random biological variation in the study population.

Inference space

Inference space is the range of organisms and environmental contexts for which the inference of an experiment is valid.

Internal validity

Internal validity refers to whether the effects observed in a study owe to manipulation of the independent variables and not some other, unknown factors.

Norm of reaction

The norm of reaction is a property of a genotype, describing how an environmental factor affects the development of the phenotype. It can be conceptualized as a function mapping expected phenotypic trait values onto environmental parameter values.

Phenotype

The phenotype of an organism is the sum of an organism's observable characteristics or traits, including its morphological, biochemical and physiological processes, behaviour and responses to external stimulation and treatments.

Phenotypic plasticity

Phenotypic plasticity describes the extent to which an organism changes its phenotype in response to environmental influences.

Random noise

Random noise (or measurement error) refers to unexplained variability in the data. It affects the variation but not the size of an experimental treatment effect.

Reproducibility

Reproducibility is the ability to produce similar results by an independent replicate experiment using the same methodology in the same or a different laboratory.

Robustness

Robustness refers to the ability of an organism to maintain a functioning phenotype under varying environmental conditions. It also refers to the stability of a response to an experimental treatment in the face of variation in environmental conditions.

3Rs principle

The guiding principles for a responsible approach to experimental animal research. They imply that a study involving the use of animals should be conducted only if the intended outcome cannot be achieved by use of no or non- sentient animals (replace), fewer animals (reduce) or procedures that are less harmful or improve animal well- being (refine).

Scientific rigor

As defined by the NIH, scientific rigor means “the strict application of the scientific method to ensure robust and unbiased experimental design, methodology, analysis, interpretation and reporting of results. This includes full transparency in reporting experimental details so that others may reproduce and extend the findings”.

Standardization

Standardization is the practice of minimizing both technical and biological variation in the study outcomes by identifying and controlling sources of variation that are believed to be putative confounders. Standardization can aim at aspects of the environment in which a study is conducted (environmental standardization), aspects of the study subjects (phenotype standardization) or aspects of how procedures and interventions are carried out and how measurements are taken (operational standardization).

Acknowledgements

The Swiss National Science Foundation (SNSF, IZSEZ0_184010) provided funding to B.V. for organising the workshop "Variation in in-vivo experiments: the norm of reaction and reproducibility". B.V., H.W. and M.K. were funded by the European Union Horizon 2020 research and innovation programme and EFPIA (Innovative Medicines Initiative, IMI 2, grant agreement No. 777364, European Quality In Preclinical Data, EQIPD). A.F. would like to thank Linnaeus University for funding. H.S. was supported by the German Research Foundation (DFG, INST 215/543-1, 396782608). I.J. was funded by the Swiss National Science Foundation (SNSF, 310030 179254).

References

1. Agassi, J. *The very idea of modern science: Francis Bacon and Robert Boyle*. vol. 298 (Springer Science & Business Media, 2012).

2. Ioannidis, J. P. A. Why most published research findings are false. *PLoS Med.* **2**, e124 (2005).

3. Open Science Collaboration. Estimating the reproducibility of psychological science. *Science* **349**, aac4716 (2015).

4. Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature* 452 (2016).

5. Munafò, M. R. *et al.* A manifesto for reproducible science. *Nat. Hum. Behav.* **1**, 21 (2017).

6. Loken, E. & Gelman, A. Measurement error and the replication crisis. *Science* **355**, 584–585 (2017).

7. Prinz, F., Schlange, T. & Asadullah, K. Believe it or not: How much can we rely on published data on potential drug targets? *Nat. Rev. Drug Discov.* **10**, 712–713 (2011).

8. Begley, C. G. & Ellis, L. M. Drug development: Raise standards for preclinical cancer research. *Nature* **483**, 531 (2012).

9. Lithgow, G. J., Driscoll, M. & Phillips, P. A long journey to reproducible results. *Nature* **548**, 387–388 (2017).

10. Collins, F. S. & Tabak, L. A. Policy: NIH plans to enhance reproducibility. *Nature* **505**, 612 (2014).

11. Freedman, L. P., Cockburn, I. M. & Simcoe, T. S. The economics of reproducibility in preclinical research. *PLoS Biol.* **13**, 1–9 (2015).

12. Goodman, S. N., Fanelli, D. & Ioannidis, J. P. A. What does research reproducibility mean? *Sci. Transl. Med.* **8**, 341ps12 (2016).

13. Forsman, A. Rethinking phenotypic plasticity and its consequences for individuals, populations and species. *Heredity* **115**, 276–284 (2015).

14. West-Eberhardt, M. J. *Developmental Plasticity and Evolution*. (Oxford University Press, 2003).

15. Stearns, S. The evolutionary significance of phenotypic plasticity. *Bioscience* **39**, 436–445 (2012).

16. Freund, J. *et al.* Emergence of individuality in genetically identical mice. *Science* **340**, 756–759 (2013).

17. Woltereck, R. Weitere experimentelle Untersuchungen über Artveränderung, speziell über das Wesen quantitativer Artunterschiede bei Daphnien. *Verh Dtsch Zool Ges* **19**, 110–172 (1909).

18. Schmalhausen, I. *Factors of Evolution; The Theory of Stabilizing Selection*. (Blakiston, 1949).

19. Hartman IV, J. L., Garvik, B. & Hartwell, L. Cell biology: Principles for the buffering of genetic variation. *Science* **291**, 1001–1004 (2001).

20. Halldorsdottir, T. & Binder, E. B. Gene × environment interactions: From molecular mechanisms to behavior. *Annu. Rev. Psychol.* **3**, 215–241 (2017).

21. Meaney, M. J. Epigenetics and the biological definition of gene x environment interactions. *Child Dev.* **81**, 41–79 (2010).

22. Cortijo, S. *et al.* Mapping the epigenetic basis of complex traits. *Science* **343**, 1145–1148 (2014).

23. Chesler, E. J., Wilson, S. G., Lariviere, W. R., Rodriguez-Zas, S. L. & Mogil, J. S. Influences of laboratory environment on behavior. *Nat. Neurosci.* **5**, 1101–1102 (2002).

24. Gururajan, A., Reif, A., Cryan, J. F. & Slattery, D. A. The future of rodent models in depression research. *Nat. Rev. Neurosci.* **20**, 686–701 (2019).

25. Beynen, A. C., Gärtner, K. & van Zutphen, L. F. M. Chapter 5: Standardization of animal experimentation. in *in Principles of laboratory animal science* (eds. Zutphen, L. F. M., Baumans, V. & Beynen, A. C.) 103–110 (Elsevier, 2003).

26. Laukens, D., Brinkman, B. M., Raes, J., De Vos, M. & Vandenabeele, P. Heterogeneity of the gut microbiome in mice: Guidelines for optimizing experimental design. *FEMS Microbiol. Rev.* **40**, 117–132 (2015).

27. Willmann, R. *et al.* Enhancing translation: Guidelines for standard pre-clinical experiments in mdx mice. *Neuromuscul. Disord.* **22**, 43–49 (2012).

28. Holmes, C., McDonald, F., Jones, M., Ozdemir, V. & Graham, J. E. Standardization and omics science: Technical and social dimensions are inseparable and demand symmetrical study. *Omics* **14**, 327–332 (2010).

29. Richter, S. H., Garner, J. P. & Würbel, H. Environmental standardization: Cure or cause of poor reproducibility in animal experiments? *Nat. Methods* **6**, 257–261 (2009).

30. Weihe, W. H. Adaptation in animal husbandry and experiment. in *Welfare and science: Proceedings of the fifth symposium of the ederation of European Laboratory Animal Science Associations, 8-11 June 1993, Brighton, UK* (London: Royal Society of Medicine Press, 1994., 1993).

31. Gur, E. & Waner, T. The variability of organ weight background data in rats. *Lab Anim.* **27**, 65–72 (1993).

32. Roe, F. J. C. Historical histopathological control data for laboratory rodents: valuable treasure or worthless trash? *Lab Anim.* **28**, 148–154 (1994).

33. Festing, M. F. Refinement and reduction through the control of variation. *Altern. to Lab. Anim.* **32**, 259–263 (2004).

34. Russell, W. M. S. & Burch, R. L. *The Principles of Humane Experimental Technique*. (1959).

35. Festing, M. F. W. Evidence should trump intuition by preferring inbred strains to outbred stocks in preclinical research. *ILAR J.* **55**, 399–404 (2014).

36. Tsai, P. P., Stelzer, H. D., Hedrich, H. J. & Hackbarth, H. Are the effects of different enrichment designs on the physiology and behaviour of DBA/2 mice consistent? *Lab Anim.* **37**, 314–327 (2003).

37. Mogil, J. S. Sex differences in pain and pain inhibition: Multiple explanations of a controversial phenomenon. *Nat. Rev. Neurosci.* **13**, 859 (2012).

38. Sorge, R. E. *et al.* Different immune cells mediate mechanical pain hypersensitivity in male and female mice. *Nat. Neurosci.* **18**, 1081 (2015).

39. Crabbe, J. C., Wahlsten, D. & Dudek, B. C. Genetics of mouse behavior: Interactions with laboratory environment. *Science* **284**, 1670–1672 (1999).

40. Loos, M. *et al.* Within-strain variation in behavior differs consistently between common inbred strains of mice. *Mamm. Genome* **26**, 348–354 (2015).

41. Prendergast, B. J., Onishi, K. G. & Zucker, I. Female mice liberated for inclusion in neuroscience and biomedical research. *Neurosci. Biobehav. Rev.* **40**, 1–5 (2014).

42. Kitano, H. Biological robustness. *Nat. Rev. Genet.* **5**, 826–837 (2004).

43. Mueller, F. S., Polesel, M., Richetto, J., Meyer, U. & Weber-Stadlbauer, U. Mouse models of maternal immune activation: Mind your caging system! *Brain Behav Immun* **73**, 643–660 (2018).

44. Kallnik, M. *et al.* Impact of IVC housing on emotionality and fear learning in male C3HeB/FeJ and C57BL/6J mice. *Mamm. Genome* **18**, 173–186 (2007).

45. Logge, W., Kingham, J. & Karl, T. Behavioural consequences of IVC cages on male and female C57BL/6J mice. *Neuroscience* **237**, 285–293 (2013).

46. Logge, W., Kingham, J. & Karl, T. Do individually ventilated cage systems generate a problem for genetic mouse model research? *Genes Brain Behav* **13**, 713–720 (2014).

47. Gur, E. *et al.* The variability of organ weight background data in rats. *Omics* **27**, 65–72 (1999).

48. Lazic, S. E. & Essioux, L. Improving basic and translational science by accounting for litter-to-litter variation in animal models. *BMC Neurosci.* **14**, 37 (2013).

49. Kimmelman, J., Mogil, J. S. & Dirnagl, U. Distinguishing between exploratory and confirmatory preclinical research will improve translation. *PLoS Biol.* **12**, e1001863 (2014).

50. Garner, J. P. 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 J.* **55**, 438–456 (2014).

51. Nosek, B. A. & Errington, T. M. Reproducibility in cancer biology: Making sense of replications. *eLife* **6**, e23383 (2017).

52. Corrigan, J. K. *et al.* A big-data approach to understanding metabolic rate and response to obesity in laboratory mice. *BioRxiv* 839076 (2019).

53. van der Staay, F. J., Arndt, S. S. & Nordquist, R. E. The standardization–generalization dilemma: A way out. *Genes, Brain Behav.* **9**, 849–855 (2010).

54. Amrhein, V., Trafimow, D. & Greenland, S. Inferential statistics as descriptive statistics: There is no replication crisis if we don't expect replication. *Am. Stat.* **73**, 262–270 (2019).

55. Servick, K. Of mice and microbes. *Science* **353**, 741–743 (2016).

56. Stappenbeck, T. S. & Virgin, H. W. Accounting for reciprocal host-microbiome interactions in experimental science. *Nature* **534**, 191–199 (2016).

57. van Driel, K. S. & Talling, J. C. Familiarity increases consistency in animal tests. *Behav. Brain Res.* **159**, 243–245 (2005).

58. Sorge, R. E. *et al.* Olfactory exposure to males, including men, causes stress and related analgesia in rodents. *Nat. Methods* **11**, 629–632 (2014).

59. Wahlsten, D. *et al.* Different data from different labs: Lessons from studies of gene-environment interaction. *J. Neurobiol.* **54**, 283–311 (2003).

60. Karp, N. A. *et al.* Impact of temporal variation on design and analysis of mouse knockout phenotyping studies. *PLoS One* **9**, e111239 (2014).

61. Fisher, R. A. *The Design of Experiments*. (Oliver and Boyd, 1935).

62. Voelkl, B. & Würbel, H. Reproducibility crisis: Are we ignoring reaction norms? *Trends Pharmacol. Sci.* **37**, 509–510 (2016).

63. Würbel, H. Behaviour and the standardization fallacy. *Nat. Genet.* **26**, 263 (2000).

64. Kafkafi, N. *et al.* Addressing reproducibility in single-laboratory phenotyping experiments. *Nat Meth* **14**, 462 (2017).

65. Richter, S. H. *et al.* Effect of population heterogenization on the reproducibility of mouse behavior: A multi-laboratory study. *PLoS One* **6**, e16461 (2011).

66. Richter, S. H., Garner, J. P., Auer, C., Kunert, J. & Würbel, H. Systematic variation improves reproducibility of animal experiments. *Nat. Methods* **7**, 167–168 (2010).

67. Voelkl, B., Vogt, L., Sena, E. S. & Würbel, H. Reproducibility of preclinical animal research improves with heterogeneity of study samples. *PLoS Biol.* **16**, e2003693 (2018).

68. Bodden, C. *et al.* Heterogenising study samples across testing time improves reproducibility of behavioural data. *Sci. Rep.* **9**, 8247 (2019).

69. Jonker, R. M., Guenther, A., Engqvist, L. & Schmoll, T. Does systematic variation improve the reproducibility of animal experiments? *Nat Meth* **10**, 373 (2013).

70. Wolfinger, R. D. Reanalysis of Richter et al.(2010) on reproducibility. *Nat Meth* **10**, 373 (2013).

71. Nelder, J. A. Statistics, science and technology. *J. R. Stat. Soc. A* **149**, 109–121 (1986).

72. Mogil, J. S. & Macleod, M. R. No publication without confirmation. *Nature* **542**, 409–411 (2017).

73. Tukey, J. W. *Exploratory Data Analysis*. (Addison-Wesley, 1977).

74. Box, G. E. P. Science and Statistics. *J. Am. Stat. Assoc.* **71**, 791–799 (1976).

75. Will, T. R. *et al.* Problems and progress regarding sex bias and omission in neuroscience research. *eNeuro* **4**, (2017).

76. Zucker, I. & Beery, A. K. Males still dominate animal studies. *Nature* **465**, 690 (2010).

77. Clayton, J. A. & Collins, F. S. NIH to balance sex in cell and animal studies. *Nature* **509**, 282–283 (2014).

78. Krzywinski, M. & Altman, N. Points of significance: Analysis of variance and blocking. *Nat. Methods* **11**, 699–700 (2014).

79. Miller, L. R. *et al.* Considering sex as a biological variable in preclinical research. *FASEB J.* **31**, 29–34 (2016).

80. Würbel, H. More than 3Rs: The importance of scientific validity for harm-benefit analysis of animal research. *Lab Anim. (NY).* **46**, 164–166 (2017).

81. Paylor, R. Questioning standardization in science. *Nat. Methods* **6**, 253–254 (2009).

82. Karp, N. A. Reproducible preclinical research—Is embracing variability the answer? *PLoS Biol.* **16**, e2005413 (2018).

83. van der Staay, F. J., Arndt, S. S. & Nordquist, R. E. Evaluation of animal models of neurobehavioral disorders. *Behav. Brain Funct.* **5**, 11 (2009).

84. Lewin, K. Frontiers in group dynamics: concept, method and reality in social science; social equilibria and social change. *Hum. Relations* **1**, 5–41 (1947).

85. Karp, N. A. & Reavey, N. Sex bias in preclinical research and an exploration of how to change the status quo. *Br. J. Pharmacol.* **176**, 4107–4118 (2019).

86. McNutt, M. Journals unite for reproducibility. *Science* **346**, 679 (2014).

87. Chia, R., Achilli, F., Festing, M. F. W. & Fisher, E. M. C. The origins and uses of mouse outbred stocks. *Nat. Genet.* **37**, 1181 (2005).

88. NC3Rs. Definitions of the 3Rs (https://www.nc3rs.org.uk/the-3rs, accessed on 2019-09-26). (2019).

89. National Institutes of Health. Consideration of sex as a biological variable in NIH-funded research (Notice No. NOT-OD-15-102). (2015).

90. OECD. OECD Guidelines for the Testing of Chemicals 408. Organization for Economic Cooperation and Development. (2018).

91. EMA. ICH guideline M3(R2) on non-clinical safety studies for the conduct of human clinical trials and marketing authorisation for pharmaceuticals. EMA/CPMP/ICH/286/1995. (2013).

92. NC3Rs. ARRIVE guidelines (https://www.nc3rs.org.uk/arrive-guidelines acessed on 2020-04-13).

93. Nature Publishing Group. Reporting checklist for life sciences articles (https://www.nature.com/documents/nr-reporting-life-sciences-research.pdf accessed on 2020-04-13).

94. Ioannidis, J. P. A., Fanelli, D., Dunne, D. D. & Goodman, S. N. Meta-research: Evaluation and improvement of research methods and practices. *PLoS Biol.* **13**, e1002264 (2015).

95. Forstmeier, W., Wagenmakers, E. J. & Parker, T. H. Detecting and avoiding likely false-positive findings – a practical guide. *Biol. Rev.* **92**, 1941–1968 (2017).

96. Jarvis, M. F. & Williams, M. Irreproducibility in preclinical biomedical research: Perceptions, uncertainties, and knowledge gaps. *Trends Pharmacol. Res.* **37**, 290–302 (2015).

97. Bishop, D. Rein in the four horsemen of irreproducibility. *Nature* **568**, 435 (2019).

98. Festing, M. F. Warning: The use of heterogeneous mice may seriously damage your research. *Neurobiol. Aging* **20**, 237–244 (1999).

99. Beck, J. A. *et al.* Genealogies of mouse inbred strains. *Nat. Genet.* **24**, 23 (2000).

100. Hsieh, L. S., Wen, J. H., Miyares, L., Lombroso, P. J. & Bordey, A. Outbred CD1 mice are as suitable as inbred C57BL/6J mice in performing social tasks. *Neurosci. Lett.* **637**, 142–147 (2017).

101. Silva, A. J. *et al.* Mutant mice and neuroscience: Recommendations concerning genetic background. *Neuron* **19**, 755–759 (1997).

102. Bogue, M. A., Churchill, G. A. & Chesler, E. J. Collaborative cross and diversity outbred data resources in the mouse phenome database. *Mamm. Genome* **26**, 511–520 (2015).

103. Tannenbaum, C., Ellis, R. P., Eyssel, F., Zou, J. & Schiebinger, L. Sex and gender analysis improves science and engineering. *Nature* **575**, 137–146 (2019).

104. Buch, T. *et al.* Benefits of a factorial design focusing on inclusion of female and male animals in one experiment. *J. Mol. Med.* **97**, 871–877 (2019).

105. Biggers, J. D. & Claringbold, P. J. Why use inbred lines? *Nature* **174**, 596 (1954).

106. Jensen, V. S., Porsgaard, T., Lykkesfeldt, J. & Hvid, H. Rodent model choice has major impact on variability of standard preclinical readouts associated with diabetes and obesity research. *Am. J. Transl. Res.* **8**, 3574 (2016).

107. Tuttle, A. H., Philip, V. M., Chesler, E. J. & Mogil, J. S. Comparing phenotypic variation between inbred and outbred mice. *Nat. Methods* **15**, 994 (2018).

108. Lerner, I. M. *Genetic Homeostasis*. (Oliver & Boyd, 1954).

109. Crusio, W. E. Inheritance of behavioral and neuroanatomical phenotypical variance: Hybrid mice are not always more stable than inbreds. *Behav. Genet.* **36**, 723–731 (2006).

110. Gingrich, J. A. & Hen, R. Commentary: The broken mouse: The role of development, plasticity and environment in the interpretation of phenotypic changes in knockout mice. *Curr. Opin. Neurobiol* **10**, 146–152 (2000).

111. Ricceri, L., Moles, A. & Crawley, J. Behavioral phenotyping of mouse models of neurodevelopmental disorders: Relevant social behavior patterns across the life span. *Behav. Brain Res.* **176**, 40–52 (2007).

112. Huang, K., Rabold, R., Schofield, B., Mitzner, W. & Tankersley, C. G. Age-dependent changes of airway and lung parenchyma in C57BL/6J mice. *J. Appl. Physiol.* **102**, 200–206 (2007).

113. Walker, C. L. *et al.* Protective effect of pregnancy for development of uterine leiomyoma. *Carcinogenesis* **22**, 2049–2052 (2001).

114. Carvalho-Freitas, M. I. R. de *et al.* Reproductive experience modifies dopaminergic function, serum levels of prolactin, and macrophage activity in female rats. *Life Sci.* **81**, 128–136 (2007).

115. Ritzel, R. M. *et al.* Multiparity improves outcomes after cerebral ischemia in female mice despite features of increased metabovascular risk. *Proc. Natl. Acad. Sci.* **114**, E5673–E5682 (2017).

116. Grimm, S. A. *et al.* DNA methylation in mice is influenced by genetics as well as sex and life experience. *Nat. Commun.* **10**, 305 (2019).

117. Richetto, J., Polesel, M. & Weber-Stadlbauer, U. Effects of light and dark phase testing on the investigation of behavioural paradigms in mice: Relevance for behavioural neuroscience. *Pharmacol. Biochem. Behav.* **178**, 19–29 (2019).

118. Sousa, N., Almeida, O. F. X. & Wotjak, C. T. A hitchhiker's guide to behavioral analysis in laboratory rodents. *Genes, Brain Behav.* **5**, 5–24 (2006).

119. Beura, L. K. *et al.* Normalizing the environment recapitulates adult human immune traits in laboratory mice. *Nature* **532**, 512 (2016).

120. Rampon, C. *et al.* Enrichment induces structural changes and recovery from nonspatial memory deficits in CA1 NMDAR1-knockout mice. *Nat. Neurosci.* **3**, 238 (2000).

121. Freedman, L. S. *et al.* Inclusion of women and minorities in clinical trials and the NIH Revitalization Act of 1993—the perspective of NIH clinical trialists. *Control. Clin. Trials* **16**, 277–285 (1995).

122. Gesensway, D. Reasons for sex-specific and gender-specific study of health topics. *Ann. Intern. Med.* **135**, 935–938 (2001).

123. Clayton, J. A. Studying both sexes: A guiding principle for biomedicine. *FASEB J.* **30**, 519–524 (2015).

124. Clayton, J. A. Applying the new SABV (sex as a biological variable) policy to research and clinical care. *Physiol. Behav.* **187**, 2–5 (2018).

125. Arnold, A. P., van Nas, A. & Lusis, A. J. Systems biology asks new questions about sex differences. *Trends Endocrinol. Metab.* **20**, 471–476 (2009).

126. Hughes, R. N. Sex does matter: Comments on the prevalence of male-only investigations of drug

effects on rodent behaviour. *Behav. Pharmacol.* **18**, 583–589 (2007).

127. Wald, C. & Wu, C. Of mice and women: The bias in animal models. *Science* **327**, 1571–1572 (2010).

128. Jazin, E. & Cahill, L. Sex differences in molecular neuroscience: From fruit flies to humans. *Nat. Rev. Neurosci.* **11**, 9 (2010).

129. Arnold, A. P. *et al.* Ischemic nitric oxide and poly (ADP-ribose) polymerase-1 in cerebral ischemia: Male toxicity, female protection. *Proc. Natl. Acad. Sci.* **20**, 565–572 (2015).

130. Beery, A. K. & Zucker, I. Sex bias in neuroscience and biomedical research. *Neurosci. Biobehav. Rev.* **35**, 565–572 (2011).

131. Klein, S. L. *et al.* Opinion: Sex inclusion in basic research drives discovery. *Proc. Natl. Acad. Sci.* **112**, 5257–5258 (2015).

132. Forsman, A. On the role of sex differences for evolution in heterogeneous and changing fitness landscapes: Insights from pygmy grasshoppers. *Philos. Trans. R. Soc. B Biol. Sci.* **373**, 20170429 (2018).

133. Yang, X. *et al.* Tissue-specific expression and regulation of sexually dimorphic genes in mice. *Genome Res.* **16**, 995–1004 (2006).

134. McCullough, L. D., Zeng, Z., Blizzard, K. K., Debchoudhury, I. & Hurn, P. D. Ischemic nitric oxide and poly (ADP-ribose) polymerase-1 in cerebral ischemia: Male toxicity, female protection. *J. Cereb. Blood Flow Metab.* **25**, 502–512 (2005).

135. Sandberg, K., Verbalis, J. G., Yosten, G. L. C. & Samson, W. K. Sex and basic science. A Title IX position. *Am. J. Physiol. Integr. Comp. Physiol.* **307**, R361–R365 (2014).

136. McCullough, L. D., McCarthy, M. M. & de Vries, G. J. NIH policy: Status quo is also costly. *Nature* **510**, 340 (2014).

137. Fields, R. D. NIH policy: Mandate goes too far. *Nature* **510**, 340 (2014).

138. Becker, J. B., Prendergast, B. J. & Liang, J. W. Female rats are not more variable than male rats: A

meta-analysis of neuroscience studies. *Biol. Sex Differ.* **7**, 34 (2016).

139. Cochran, W. G. & Cox, G. M. *Experimental Design*. (John Wiley and Sons, 1957).