PLOS BIOLOGY

ESSAY

Finding the right power balance: Better study design and collaboration can reduce dependence on statistical power

Shinichi Nakagawa[1,2]*, **Malgorzata Lagisz**[1,2], **Yefeng Yang**[1], **Szymon M. Drobniak**[1,3]

1 Evolution & Ecology Research Centre and School of Biological, Earth and Environmental Sciences, University of New South Wales, Sydney, Australia, **2** Theoretical Sciences Visiting Program, Okinawa Institute of Science and Technology Graduate University, Onna, Japan, **3** Institute of Environmental Sciences, Jagiellonian University, Kraków, Poland

* s.nakagawa@unsw.edu.au

Citation: Nakagawa S, Lagisz M, Yang Y, Drobniak SM (2024) Finding the right power balance: Better study design and collaboration can reduce dependence on statistical power. PLoS Biol 22(1): e3002423. https://doi.org/10.1371/journal.pbio.3002423

Published: January 8, 2024

Funding: SN and ML were supported by the Australian Research Council (ARC) Discovery Project Grant awarded (DP210100812), and SMD was supported by the ARC Discovery Early Career Award (DE180100202). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abbreviations: AHARP, as high as reasonably practicable; ALARP, as low as reasonably practicable; EDI, equitable, diverse, and inclusive; FAIR, findable, accessible, interoperable, and reusable; GLMM, generalized linear mixed(-effects) model; GWAS, genome-wide association study; HARKing, hypothesizing after results are known;

Abstract

Power analysis currently dominates sample size determination for experiments, particularly in grant and ethics applications. Yet, this focus could paradoxically result in suboptimal study design because publication biases towards studies with the largest effects can lead to the overestimation of effect sizes. In this Essay, we propose a paradigm shift towards better study designs that focus less on statistical power. We also advocate for (pre)registration and obligatory reporting of all results (regardless of statistical significance), better facilitation of team science and multi-institutional collaboration that incorporates heterogenization, and the use of prospective and living meta-analyses to generate generalizable results. Such changes could make science more effective and, potentially, more equitable, helping to cultivate better collaborations.

Introduction

Given how much scientific progress has been made and how it is accelerating, it feels paradoxical to discover that >80% of research is potentially "wasted." Two independent estimates from the fields of medicine and ecology confirm that this is the case [1,2]. The 2 primary sources of such waste are suboptimal study design and selective publication and reporting (the latter we refer to collectively as publication bias) [1–3]. Null hypothesis significance testing (NHST; Box 1), or more precisely, the misuse of NHST, may be the main culprit behind the issue of such publication bias because it makes the continuous nature of evidence artificially binary by using the threshold of p-values ($\alpha = 0.05$) [4,5]. NHST facilitates not only selective publication and reporting but also p-hacking, HARKing (hypothesizing after results are known), and other types of what are known as "questionable research practices" [6,7]. Such misuses of NHST have been recently linked to massive failures to replicate published studies in many fields, the so-called "replication crisis" [8–10]. Indeed, researchers have been criticizing NHST for at least three-quarters of a century [11–13].

After decades of controversies and criticisms on NHST and p-values, it is somewhat surprising that concepts of statistical power and power analysis still seem to enjoy freedom from

NHST, null hypothesis significance testing; RCT, randomized controlled trial.

Box 1. Glossary

Null hypothesis significance testing (NHST)

In this framework, a null hypothesis is assumed (usually zero effect) for an intervention or phenomenon. After an experiment or observation, if the inferential statistic obtains a *p*-value of less than (or equal to) 0.05, the null hypothesis is rejected and the alternative hypothesis of nonzero effect is accepted (i.e., statistically significant or positive results). If a *p*-value higher than 0.05 is obtained, the null hypothesis is retained (i.e., nonsignificant or negative results).

p-hacking

The NHST framework incentivizes *p*-values of less than or equal to 0.05. Therefore, arbitrary analytical decisions are often made to reach statistically significant results. For example, researchers might keep fitting different predictors (independent variables) to their statistical models until they produce a statistically significant result. *p*-hacking is one of the most common questionable research practices.

HARKing

The term represents an abbreviation for hypothesizing after results are known (HARKing). HARKing is a questionable research practice in which researchers generate a hypothesis to fit their known results so that they get positive results, which are easier to publish than negative results. A hypothesis should be created a priori.

Linear mixed modeling

It encompasses a group of statistical models with fixed effects and random effects, therefore often referred to as mixed-effects models. The model estimates regression coefficients for fixed effects, while it estimates variance components for random effects. The term "linear mixed(-effects) models" often indicate models assuming the Gaussian (normal) error structure but can include models with non-Gaussian errors, such as Poisson and binomial errors, which are often referred to as generalized linear mixed(-effects) models (GLMMs).

similar condemnations and disapprovals [14–16]. Power analysis is often used to determine a sufficient sample size necessary for *statistical significance*, thus fully endorsing NHST (see Box 2). Yet, at the same time, power analysis—when used correctly—improves study design by providing a sample size that gives more precise estimates (e.g., smaller standard error) [11,17]. Therefore, 2 powerful gatekeepers of academia, grant agencies, and ethics committees endorse and (at least indirectly) recommend power analysis for sample size calculations [18]. Their argument is that it is unethical and a waste of money to conduct underpowered or overpowered studies, leading to the recommendation of the nominal 80% statistical power. This argument seems undoubtedly true, especially for human trials [19,20]. Of relevance, researchers are often criticized for a lack of sufficient statistical power (or power analysis), not only in grant applications but also in scientific manuscripts, despite planning or doing the best they can within the constraints of time, finance, and other logistics.

Box 2. Power analysis and related concepts

Power analysis involves 4 parameters: statistical power, which is 1 minus a Type II error rate ($1-\beta$), often set to be 0.80; a Type I error rate, also known as significance level, α, usually fixed at 0.05; sample size, N; and standardized effect size, $E[\theta]/\sqrt{Var[\theta]}$, where θ is the effect size of interest and its population average ($E[\theta]$) and variance ($Var[\theta]$). If we know 3 of these 4, we can calculate the fourth unknown parameter.

Power analysis usually requires some estimates of standardized effect size (note that standardized mean difference d is an example of standardized effect size [21]). However, it is often challenging to obtain a good estimate, and published estimates are likely to be inflated [14,22,23]. It is interesting to note, when $E[\theta]/\sqrt{Var[\theta]}$ is examined, that there are 2 routes to having a large standardized effect size: either via having a large estimate of the population effect $E[\theta]$ or by having a small estimate of population variance $Var[\theta]$. Indeed, in the vicious cycle of power analysis (see section on "The vicious cycle of publication bias and power analysis"), both are simultaneously happening, boosting the magnitude of the standardized effect size.

Assuming $\alpha = 0.05$, $(1-\beta) = 0.8$, and the d values are as in the main text (e.g., $d = 0.125$), one can use the following formula to approximate the sample size required for 1 group of 2 independent sample groups [24]:

$$N = 16\frac{Var[\theta]}{E[\theta]^2} = 16\frac{1}{d^2}.$$

In **S1 Supporting Information**, we provide an R script where we calculate the sample sizes used in the examples provided. Note that the above formula is incorrect for the interaction effect (e.g., sex difference in a treatment effect), as it involves 4 groups rather than 2, so in that scenario, one needs to use 32 instead of 16.

In this Essay, we challenge the premise that 80% statistical power is necessary for addressing many basic research questions (where a realistic study will almost always be underpowered yet worthwhile to conduct). We discuss how the misuse of power analysis contributes to research waste and the replication crisis in a nontrivial way and argue that undue focus on statistical power, similar to that on p-values, could counterintuitively encourage scientists to choose non-optimal designs rather than improve study design. From the viewpoint of generalizability, we suggest that a set of several low-powered studies could be better than one high-powered study, even when the combined sample sizes are comparable in both scenarios [25,26]. Importantly, we discuss a series of potential alternatives and supplements to power analysis, which researchers and gatekeepers can implement. Our proposed paradigm shift can potentially improve science and its equity simultaneously by making science more collaborative.

The vicious cycle of publication bias and power analysis

As already mentioned, one of the underlying causes of the replication crisis is publication bias, which is related to the filtering effect of NHST, causing an exaggeration of scientific evidence in terms of published effect sizes. Indeed, a series of large replication efforts have repeatedly shown that replication studies usually obtain much smaller effect sizes (e.g., 50% smaller [27])

than the original studies they sought to replicate [28–32]. In addition, recent meta-research studies have confirmed that inflated effect estimates in the published literature are common in many fields, including psychology, economics, ecology, and medicine [33–39]. For example, according to a meta-analysis of global change biology experiments that accounted for publication bias [37], a statistically significant effect reported in the literature is, on average, 2 to 3 times larger than a "true" effect. Furthermore, an average experiment in that field was severely underpowered (<40%) [35]. Therefore, published experiments often have small sample sizes, yet surprisingly large effects. The situation may be even worse for human randomized controlled trials (RCTs). A study found the median power of 23,551 RCTs to be only approximately 13% [23], probably because sample sizes were determined on the basis of inflated effects that had been previously reported.

When deciding on a sample size for a new study, the most common method is to use a closely related published study or studies to generate an effect size estimate on which to base the power analysis [40]. But we know that published significant effects are inflated because of publication bias [14,22,23]. The consequence of using an overestimated effect in the power analysis is a sample size estimate that is far smaller than what is actually needed to "detect" a true effect [41]. Yet, sometimes, a value of $p < 0.05$ can be achieved by chance in this scenario, leading to the publication of yet another "inflated" effect, keeping this unfortunate and vicious cycle of power analysis going (also referred to as the winner's curse [23,42,43]; see Fig 1).

We would argue that this cycle substantially contributes to research waste and the replication crisis. Inadvertently, grant agencies and ethics committees hold a key role in perpetuating this vicious cycle, as they endorse (and often require) power analysis. Some readers, particularly statisticians, may argue that this is not the fault of power analysis (and NHST) but of researchers who misuse it. However, given the prevalence of low statistical power in many studies, including RCTs [23], we believe that a critical rethink of how power analysis should be used or recommended is necessary.

Two opposing forces

The current incentive structure and requirements of academia pull researchers in 2 completely opposite and incompatible directions: towards studies with small sample sizes (hereafter, small studies) and towards studies with large sample sizes (large studies). The prevalence of low-powered studies suggests that forces encouraging small studies are very strong. Research operates within the parameters of limited resources and time and a complex landscape of ethical regulation, all creating a huge incentive to conduct less costly experiments with small sample sizes. Such small studies will appear to have enough statistical power when designed under the expectation of a large effect size estimate, and researchers have no trouble finding such large, yet inflated, effect estimates in the literature. Resorting to meta-analytical estimates does not alleviate the issue. Although often more conservative (e.g., through active retrieval of unpublished estimates), such estimates are not free from publication bias and effect size inflation. Logistics aside, grant agencies usually appreciate the "value for money" offered by small studies, and ethics committees often prefer smaller to larger studies, thereby enabling researchers to maintain the vicious cycle together with grant and ethics boards.

By contrast, forces that encourage larger studies are present but often neglected. A study based on 13,322,754 abstracts from PubMed demonstrated that effect sizes declined between 1990 and 2015, while the frequency of statistically significant results increased, indicating the sample sizes of studies increased over the same period [44]. Academia pursues novelty, and such pursuits usually lead to the testing of more complex and subtle effects because the most obvious and large effects have usually already been discovered [14,45]. A case in point is gene–

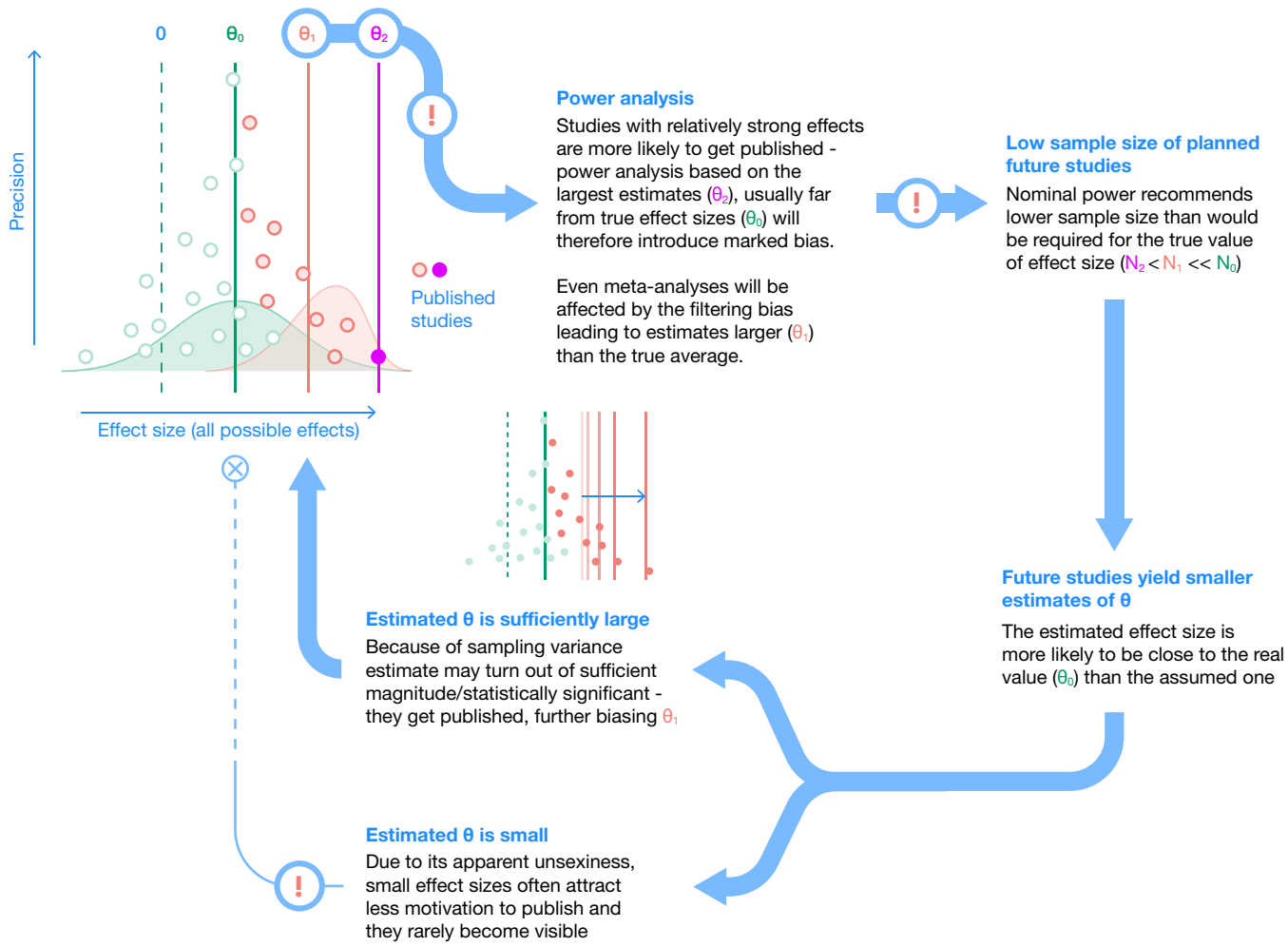

Fig 1. The vicious cycle of power analysis and publication bias. An example of how effect size θ can be inflated via selective publication and how power analysis, in its current use, can encourage this cycle to continue.

https://doi.org/10.1371/journal.pbio.3002423.g001

trait association studies where, in the early years, researchers were able to find genes with large effects, while more recently, such a discovery is rare; indeed, in recent years, most genome-wide association studies (GWASs) find many genes with small yet important effects [46]. It seems that most researchers nowadays are interested in research topics where the "true" effect is relatively small. Recent large-scale replication efforts have revealed that even effects believed to be large and general are usually small and too subtle to be useful or are even nonexistent, particularly in psychology [27–32]. It requires at least hundreds, if not thousands, of subjects to conduct an experiment that finds a significant yet small effect, which may be out of reach for many researchers.

Notably, our discussion has so far focused only on the main effect size in a study. To study interaction effects (e.g., sex differences in the treatment effect [47]), an 8-times larger sample size will be needed. This is the case when the interaction is the same magnitude as the main effect. A 16-times larger sample size is needed if it is assumed that the interaction is half of the main effect, which is more realistic [48]. Indeed, novel and important questions may often reside in interaction effects [49], which are usually smaller than the main effect size. Therefore, the implicit and explicit requirements of 80% power could stop researchers from exploring

this frontier of knowledge. Such small effects relate to the idea that researchers should use the smallest effect size of interest for power analysis [50,51] (for an alternative, see [52]); however, using the smallest effect of interest often requires a larger study, which consequently requires more funding to perform (see **S1 Supporting Information**).

Of relevance, requirements for relatively large sample sizes (e.g., $N = 100$) would exclude many vertebrate researchers, particularly conservation biologists, from conducting their studies [53]. Furthermore, although labs that can afford large studies might manage to find a small, yet important effect, replicability and generalizability are far from guaranteed. If results are to be generalizable, experiments should include heterogenization, for example, by including different strains of animals and a range of environmental conditions [26,54,55] (Box 3). Thorough heterogenization necessarily increases within-subject variability, as it covers the landscape of different effect size magnitudes and variations (see Fig 2 in Box 3). So does heterogenization require an increase in sample size to maintain statistical power? Imagine that a researcher wants to heterogenize their 40 mice with regard to their strains. If they could get 20 different strains and create a complete block design by creating 20 blocks (i.e., each stain is assigned in both control and treatment groups), then they will not need to increase the sample

Box 3. The importance of variability due to plasticity and heterogenization

Variance observed in measured outcomes of empirical studies comes not only from between-individual variance and sampling error but also from environmental variance generated by the dependence of traits on external environmental variables (i.e., on the shape of a trait's reaction norm [56–58]). Ignoring the reaction norm and forcing empirical studies (controlled experiments in particular) to eliminate all sources of environmental variation deemed "irrelevant" leads to increasingly irreplicable outcomes that simply explore different regions of a reaction norm mapping function [56,59] (Fig 2). Individual empirical studies focus on very specific environmental conditions to reduce unwanted variation in measured traits and amplify the expected differences (i.e., different points on the x axis in Fig 2). However, doing so in the presence of any meaningful relationship between the environment (x) and the measured trait (y) generates apparent discordance in observed phenotypes generated purely by their environmental plasticity.

If too much focus is given to maximizing statistical power (or precision), this process leads to an interesting paradox [56]. To measure traits as precisely as possible, individual studies generate more specific, nonoverlapping outcomes that hamper the reproducibility of key results. The solution is to rely less on a specific study and more on the comprehensive exploration of the underlying gradient of environmental variability [26,54,60]. In fact, less precise (e.g., lower powered) studies could paradoxically improve reproducibility as they generate outcomes that are not in conflict (note the overlap of the less precise blue density with 2 more precise and disconnected red and green densities on the y axis in Fig 2). Therefore, paying less attention to power analysis is only part of the solution. When coupled with a wider shift of the empirical paradigm (e.g., through heterogenization to represent whole ranges of underlying environmental and/or genetic variation in planned experiments [54,61,62]), we can move closer to resolving the ongoing reproducibility crisis.

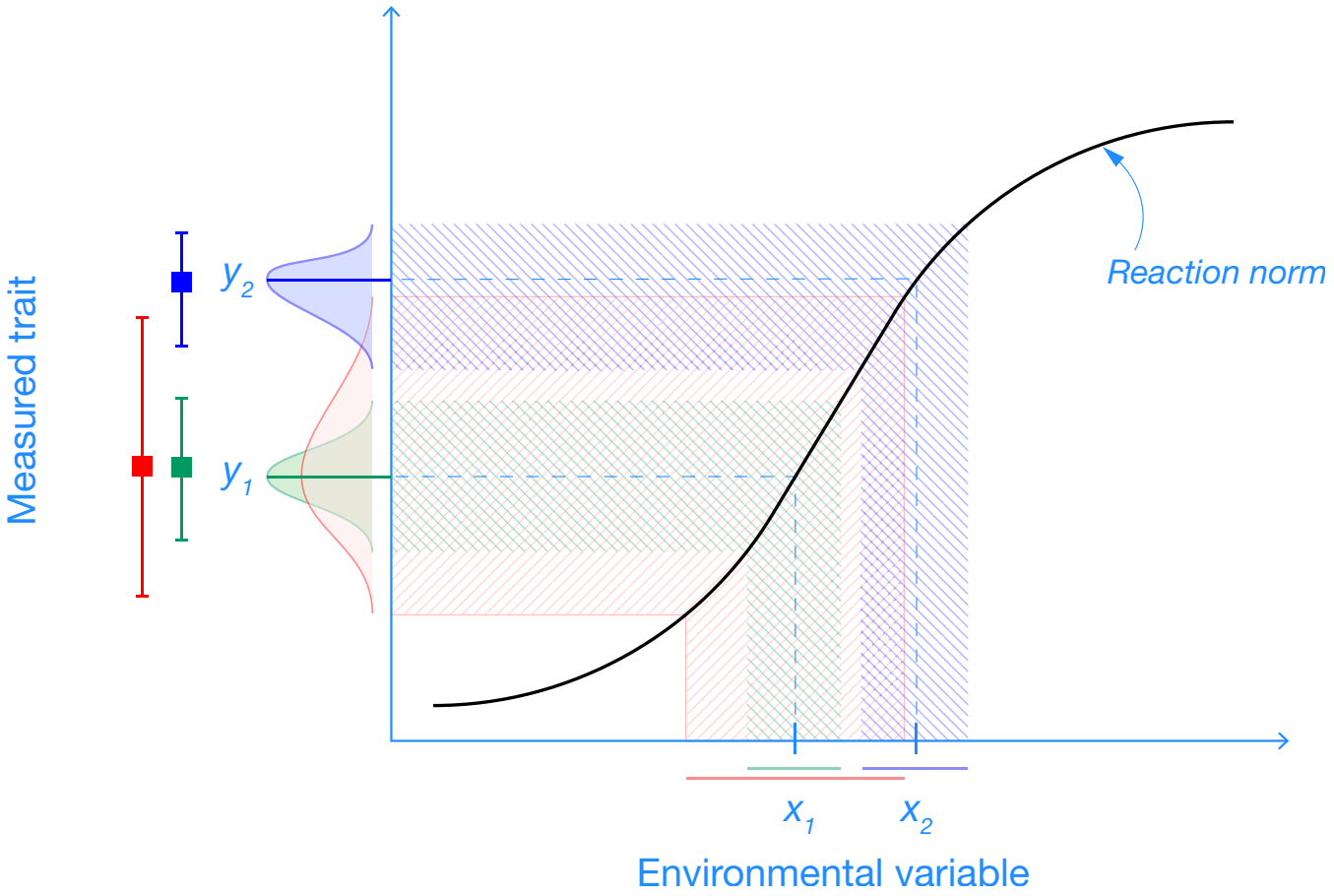

Fig 2. Plasticity of a trait in relation to an environmental variable. Traits are expressed differently (y_1 and y_2) depending on environmental conditions (x_1 and x_2). Therefore, excessive standardizations (of environments) will lead to unreplicable results. See Box 3 for the details of differently colored parts.

https://doi.org/10.1371/journal.pbio.3002423.g002

size [54]. Yet, in reality, they are likely to get only 5 strains, creating replicates per control/treatment within the 5 blocks. In such a case, they do need to increase their sample size because mice within blocks (the same strains) are more similar to each other (i.e., not independent [54]). Taken together, how can the vicious cycle be escaped from, without also requiring a large sample size for many research questions? We argue that we must find and achieve a happy medium.

Better study design with less emphasis on power

The current focus on power will not help resolve the issue of 2 different forces acting on researchers. The best thing to do, therefore, is to shift attention to generating a better study design without worrying too much about reaching the nominal statistical power of 80% (apart from situations where large effects are expected, such as with pharmacological and toxicological interventions). We suggest using the AHARP (as high as reasonably practicable) principle, mirroring the ALARP (as low as reasonably practicable) principle, which is used in health and safety [63]. The AHARP principle assumes that it is often impossible to achieve enough power in a study when small effects and generalizations are considered. This principle aims to attain the best possible power or precision for a study within the constraints of budget and resources so that everybody (no matter their financial situation) can participate in research activities.

Such a principle could mean that studies can have a relatively small sample size and be underpowered (e.g., $N < 100$). It is already known that small studies produce imprecise results [33], yet it is important to realize that, collectively, small studies themselves are unbiased; in other words, averaging results from many small studies could provide an accurate estimate of a "true" effect (see **S1 Supporting Information**). It is the filtering effect of the publication process on the basis of statistical significance (Fig 1) or other related criteria that produce exaggerated effect sizes, thereby making science unreliable.

Importantly, when we say "less emphasis on power" or emphasize the AHARP principle, this does not mean we think that the power (or precision) of studies should be ignored altogether. We only suggest that well-conceptualized study plans should not be cast aside because they fail to reach the expected >80% power. Thus, the AHARP principle does not equate to "free-for-all" research, and we would remind researchers that there are other aspects of study design to consider beyond just increasing sample size to improve study power and precision [64]. However, covering all aspects of study design is beyond the scope of this Essay (for further discussions, see [65,66]).

In many cases, statistical power can be improved by explicitly incorporating correlated structures between treatment and control groups, compared to using independent subjects alone (e.g., using sibling pairs as a complete block design; see **S1 Supporting Information**). By contrast, nested or hierarchical structures (e.g., siblings within mothers or animals within strains) could reduce power when such structures are statistically accounted for (if such structures are not accounted for, it is known as pseudoreplication [67–69]). Such correlated, nested, or hierarchical structures can be explicitly modelled using a (generalized) linear mixed modelling approach [70,71]. However, such complex designs pose difficulties for estimating precision and conducting power analysis. This is because the conventional algebraic formulas (Box 2) cannot be used to estimate the necessary sample size, so simulation must be used instead, which can become very complex [72–74]. One of the reasons for the difficulties is the necessity of knowing how correlated the data from a cluster is (e.g., how similar pups from the same mother are for a given measurement; see **S1 Supporting Information**). Nevertheless, researchers should be aware of the uses of correlated samples and that modelling correctly can provide a more precise and higher-powered design.

Researchers can also try to increase the precision of their measurements. For example, it is becoming increasingly easier to measure behavioral traits more precisely with AI-assisted video recording analyses [75]. Although not easy and potentially time-consuming, researchers could choose to optimize their study design, including improving their sampling strategies and using more precise measurement techniques, rather than relying upon power analysis, the correct implementation of which is often very difficult. Once they have their "best" sampling design regardless of its power, researchers may want to conduct a "design analysis," which is defined as "a set of statistical calculations about what could happen under hypothetical replication of a study—that focuses on estimates and uncertainties rather than on statistical significance" [14] (see also [10]). The main part of the design analysis is calculating Type S (sign) error (the probability of getting the sign wrong when a result is statistically significant) and Type M (magnitude) error (the degree to which an effect is overestimated when significant) [14,22]. Type S and Type M errors are also defined in terms of statistical significance, but these 2 types of errors focus on estimates rather than significance [14].

To make our position clear, we think the concepts of statistical significance and power, along with *p*-values and power analysis, are important for navigating the scientific literature and, when used correctly, can be useful [76,77]. However, we feel that grant agencies (including grant assessors) and ethics committees should be satisfied if researchers have done due diligence when coming up with the best study design. If researchers can report their study

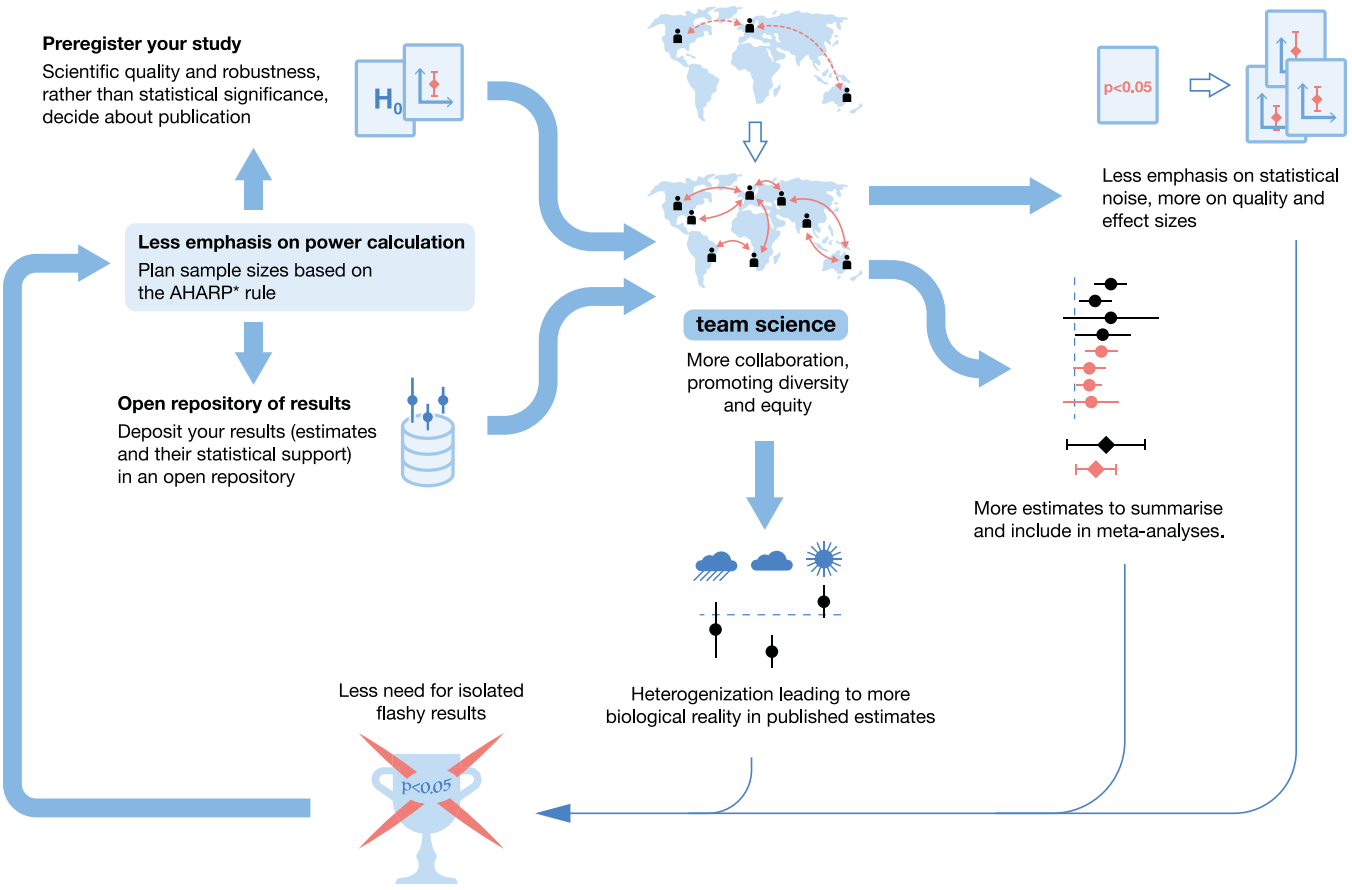

Fig 3. The virtuous cycle of research. A visualization of how our proposed paradigm shift could start a virtuous cycle that empowers researchers and better science.

https://doi.org/10.1371/journal.pbio.3002423.g003

design's Type S and Type M error rates, we believe this would provide a better benchmark for a proposed empirical project than conventional power analysis: We would even encourage researchers to report statistical power along with Type S and Type M error rates.

From vicious cycle to virtuous cycle

We believe that grant and ethics bodies should not always expect researchers to determine sample size via power analysis. As argued above, such usage of power analysis might influence researchers to choose suboptimal study designs and could maintain the vicious cycle of biased research findings and research waste (Fig 1). Instead, researchers, grant agencies, and ethics boards could be working together to turn the vicious cycle into a virtuous cycle (Fig 3).

Registration and full reporting

Power analysis, used wrongly, could eliminate interesting research ideas that could otherwise, in accumulation, contribute to a field. Instead, grant agencies could ask researchers to (pre) register their funded and approved studies (note that the terms "registration" and "preregistration" are used interchangeably for the same process [78]) and publish their work regardless of the statistical significance of the results (Fig 3). We propose that funders and journals team up

to ensure that all registered studies are published, regardless of their results. According to some estimates, more than 50% of studies remain unpublished, mainly because the results did not reach statistical significance [2,79]. Registration, along with registered reports, can partially mitigate this issue [80]. Relatedly, there are novel ways of disseminating research, such as Octopus and ResearchEquals, both of which allow different components of research to be published in a separate yet modular manner (e.g., hypothesis, method, result, code).

Unfortunately, we do not think that registration and these related innovations will be the main solution for publishing negative results. For years, scientists have repeatedly argued, with little effect, that a study needs to be published regardless of statistical significance, yet it seems that much research remains unpublished [2,79]. This is not surprising because proper incentives for doing so are not yet in place. Therefore, we propose that a free repository of statistically nonsignificant results (or all results) be created, preferably associated with study registration. In this repository, one could fill in study results using a template in a short amount of time, making the data findable, accessible, interoperable, and reusable (FAIR) [81]. Setting up such a repository, and mandating its use, is exactly what grant agencies and ethics committees could be doing. Archiving nonstatistically significant results is essential because results from well-designed studies are unbiased regardless of statistical power or how small a study was. Such a repository would enable the community to access the results of all relevant studies for later syntheses. Reporting results to registries is mandatory for some medical RCTs, although there seem to be some issues with compliance [82,83]. Grant agencies and ethics committees could certainly help fix such compliance issues [84].

Collaboration to improve reproducibility and equity

Pluralism and diversity make science better [85,86]. In addition to needing greater pluralism, we need to realize that what one study can achieve is limited, however powerful, well-designed, and expensive such a study might be [25,87,88]. Grant agencies and ethics committees therefore have an important role in fostering and supporting collaboration for multiple studies (Fig 3).

If grant agencies and ethics committees allowed AHARP study designs, science could move towards becoming more equitable, diverse, and inclusive (EDI) [86,89]. For many emerging questions where large effects are not expected, only those with sufficient funding are able to conduct the large experiments that power analysis would demand. However, being inclusive of any studies, regardless of their power, would encourage more research from different institutions across the globe. Of importance, a simulation study indicates that even well-funded laboratories should consider conducting several low-powered studies (e.g., 30% power) rather than a single high-powered study (80% power; note that the latter is approximately 4 times larger than the former [25]). This is because when the effect of interest has a realistic amount of heterogeneity (e.g., due to meaningful temporal and locational variation), a single high-powered study has a higher Type I error rate than an aggregation of several low-powered studies, which can better accommodate heterogeneity [25]. Therefore, even the well-funded would do well to collaborate with others at different institutions to make their experimental results more robust and in line with the idea of heterogenization (Box 3). Such designs not only improve the overall power of estimates but also make them more biologically relevant and generalizable. Grant agencies, along with ethics committees, could encourage and specifically fund multi-institutional experiments, through which they could provide more opportunities to researchers from traditionally marginalized groups, spreading EDI in science [86,89] (for a related example of when and how such an experiment could be funded, see [90]). Such a multi-institutional experiment, combined with a later synthesis, can be seen as a "prospective" meta-analysis [91].

Indeed, this type of synthesis is exactly what big team science projects have done and are trying to do. In recent years, CERN-style, big team science projects have emerged and spread across many fields [92]. Examples include ManyBabies [93], the Reproducibility Project: Cancer Biology [30], SPI-Birds [94], and the Nutrient Network [95], (see also [96] for an example of how citizen science can be harnessed to increase statistical power and precision). Such team science projects form a collaborative community across several institutions to conduct a prospective meta-analysis, which resolves the post hoc nature of traditional meta-analyses. Not surprisingly, many post hoc meta-analytic estimates are also much larger than those from multilaboratory replication efforts [32] (e.g., Many Labs [97,98]). This result indicates that meta-analytic means are often overestimated, although bias-corrections of meta-analytic mean estimates are possible and can be effective [99]. Therefore, we propose a shift from traditional to prospective meta-analyses.

Big team science projects are able to do more than just produce a prospective meta-analysis because of the communities they create. Such communities can organize a meta-analysis to be continuously updated (i.e., a living synthesis) [100], which has recently been described as an "open synthesis community" [101]. Notably, team science is not without its problems; for example, there are concerns regarding how to fairly credit each scientist involved and whether team science could increase inequity rather than decrease it [92,102,103]). But this is where grant agencies could intervene to introduce new criteria for recognizing scientific contributions and make sure large collaborative efforts, which they fund, address EDI fully [87]. Regardless, it will require coordination among researchers, funders, institutions, and other relevant committees and organizations (e.g., learned societies) to make team scientific activities easier and fair [104].

Conclusions

We began this article by referring to 2 major causes of "research waste": suboptimal study design and publication bias (selective publication and reporting). We have argued that, although power analysis helps study design in theory, paying less attention to statistical power may improve study design in practice, just like paying less attention to statistical significance (threshold p-values) could alleviate the issue of publication bias. Hopefully, we have convinced many, especially those on grant and ethics committees, that it is time for a paradigm shift in our approach to research. We must encourage better study designs with less focus on power; (pre)registration and full publication of all data; team science or multi-institutional collaborations that allow realistic incorporation of heterogenization; and prospective and living meta-analyses to reach generalizable results. By adopting those changes, we can break out of the vicious cycle into the virtuous cycle (Fig 3). In such a virtuous cycle, less emphasis on statistical power could start and maintain a more collaborative, equitable, and diverse scientific environment, where both underestimates and overestimates are welcome and integrated to achieve an estimate closer to a "true" effect. To get there, we need to find the right "power" balance.

Supporting information

S1 Supporting Information. An HTML file containing 3 sections: Section 1, a fictitious story providing different experimental scenarios using mice; Section 2, calculating statistical power under the scenarios introduced under Section 2; and Section 3, showing how small low-powered studies can be aggregated via a meta-analysis.
(HTML)

Acknowledgments

We thank Diego Barneche and James McGree for their comments on an earlier version of the manuscript. A part of the writing was conducted while visiting the Okinawa Institute of Science and Technology (OIST) through the Theoretical Sciences Visiting Program (TSVP) to SN.

Author Contributions

Conceptualization: Shinichi Nakagawa, Malgorzata Lagisz, Yefeng Yang, Szymon M. Drobniak.

Formal analysis: Shinichi Nakagawa, Yefeng Yang, Szymon M. Drobniak.

Funding acquisition: Shinichi Nakagawa, Malgorzata Lagisz, Szymon M. Drobniak.

Investigation: Shinichi Nakagawa, Yefeng Yang, Szymon M. Drobniak.

Methodology: Shinichi Nakagawa, Yefeng Yang, Szymon M. Drobniak.

Project administration: Shinichi Nakagawa.

Visualization: Malgorzata Lagisz, Szymon M. Drobniak.

Writing – original draft: Shinichi Nakagawa, Szymon M. Drobniak.

Writing – review & editing: Shinichi Nakagawa, Malgorzata Lagisz, Yefeng Yang, Szymon M. Drobniak.

References

1. Chalmers I, Glasziou P. Avoidable waste in the production and reporting of research evidence. Lancet. 2009; 374(9683):86–89. https://doi.org/10.1016/S0140-6736(09)60329-9 WOS:000267768800037. PMID: 19525005
2. Purgar M, Klanjscek T, Culina A. Quantifying research waste in ecology. Nat Ecol Evol. 2022; 6 (9):1390–1397. https://doi.org/10.1038/s41559-022-01820-0 WOS:000828437000001. PMID: 35864230
3. Ioannidis JPA. Why most published research findings are false. PLoS Med. 2005; 2(8):696–701. https://doi.org/10.1371/journal.pmed.0020124 WOS:000231676900008. PMID: 16060722
4. Amrhein V, Greenland S, McShane B. Scientists rise up against statistical significance. Nature. 2019; 567(7748):305–307. Epub 2019/03/22. https://doi.org/10.1038/d41586-019-00857-9 PMID: 30894741.
5. Wasserstein RL, Schirm AL, Lazar NA. Moving to a World Beyond "p < 0.05". Am Stat. 2019; 73:1–19. https://doi.org/10.1080/00031305.2019.1583913 WOS:000462083800001.
6. John LK, Loewenstein G, Prelec D. Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. Psychol Sci. 2012; 23(5):524–532. https://doi.org/10.1177/0956797611430953 WOS:000314464500014. PMID: 22508865
7. Fraser H, Parker T, Nakagawa S, Barnett A, Fidler F. Questionable research practices in ecology and evolution. PLoS ONE. 2018; 13(7). https://doi.org/10.1371/journal.pone.0200303 WOS:000438829800017. PMID: 30011289
8. Baker M. 1,500 scientists lift the lid on reproducibility. Nature. 2016; 533(7604):452–454. Epub 2016/05/27. https://doi.org/10.1038/533452a PMID: 27225100.
9. Amrhein V, Trafinnow D, Greenland S. Inferential Statistics as Descriptive Statistics: There Is No Replication Crisis if We Don't Expect Replication. Am Stat. 2019; 73:262–270. https://doi.org/10.1080/00031305.2018.1543137 WOS:000462083800029.
10. Greenland S. Analysis goals, error-cost sensitivity, and analysis hacking: Essential considerations in hypothesis testing and multiple comparisons. Paediatr Perinat Ep. 2021; 35(1):8–23. https://doi.org/10.1111/ppe.12711 WOS:000594980400001. PMID: 33269490

11. Nakagawa S, Cuthill IC. Effect size, confidence interval and statistical significance: a practical guide for biologists. Biol Rev. 2007; 82(4):591–605. https://doi.org/10.1111/j.1469-185X.2007.00027.x ISI:000250251600004. PMID: 17944619

12. Nickerson RS. Null hypothesis significance testing: A review of an old and continuing controversy. Psychol Methods. 2000; 5(2):241–301. https://doi.org/10.1037/1082-989x.5.2.241 WOS:000087775500007. PMID: 10937333

13. Rozeboom WW. The Fallacy of the Null-Hypothesis Significance Test. Psychol Bull. 1960; 57(5):416–428. https://doi.org/10.1037/h0042040 WOS:A1960WD17200004. PMID: 13744252

14. Gelman A, Carlin J. Beyond Power Calculations: Assessing Type S (Sign) and Type M (Magnitude) Errors. Perspect Psychol Sci. 2014; 9(6):641–651. https://doi.org/10.1177/1745691614551642 WOS:000345305100006. PMID: 26186114

15. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol. 2016; 31(4):337–350. https://doi.org/10.1007/s10654-016-0149-3 WOS:000376675000002. PMID: 27209009

16. Rothman KJ, Greenland S. Planning Study Size Based on Precision Rather Than Power. Epidemiology. 2018; 29(5):599–603. https://doi.org/10.1097/EDE.0000000000000876 WOS:000441143500009. PMID: 29912015

17. Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, et al. Confidence and precision increase with high statistical power. Nat Rev Neurosci. 2013; 14(8). https://doi.org/10.1038/nrn3475-c4 WOS:000322002100015. PMID: 23820778

18. Knapp TR. The overemphasis on power analysis. Nurs Res. 1996; 45(6):379–381. https://doi.org/10.1097/00006199-199611000-00018 WOS:A1996VX70200018. PMID: 8941316

19. Halpern SD, Karlawish JHT, Berlin JA. The continuing unethical conduct of underpowered clinical trials. J Am Med Assoc. 2002; 288(3):358–362. https://doi.org/10.1001/jama.288.3.358 WOS:000176807000028. PMID: 12117401

20. Celik S, Yazici Y, Yazici H. Are sample sizes of randomized clinical trials in rheumatoid arthritis too large? Eur J Clin Invest. 2014; 44(11):1034–1044. https://doi.org/10.1111/eci.12337 WOS:000344525600002. PMID: 25207845

21. Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Hillsdale, N.J.: L. Erlbaum Associates; 1988.

22. Gelman A, Tuerlinckx FA. Type S error rates for classical and Bayesian single and multiple comparison procedures. Computation Stat. 2000; 15(3):373–390. https://doi.org/10.1007/s001800000040 WOS:000089908000004.

23. van Zwet E, Schwab S, Greenland S. Addressing exaggeration of effects from single RCTs. Significance. 2021; 18(6):16–21.

24. Lehr R. 16 S-Squared over D-Squared—a Relation for Crude Sample-Size Estimates. Stat Med. 1992; 11(8):1099–1102. https://doi.org/10.1002/sim.4780110811 WOS:A1992JB42900010. PMID: 1496197

25. IntHout J, Ioannidis JPA, Borm GF. Obtaining evidence by a single well-powered trial or several modestly powered trials. Stat Methods Med Res. 2016; 25(2):538–552. https://doi.org/10.1177/0962280212461098 WOS:000374792800003. PMID: 23070590

26. Voelkl B, Vogt L, Sena ES, Wurbel H. Reproducibility of preclinical animal research improves with heterogeneity of study samples. PLoS Biol. 2018; 16(2). https://doi.org/10.1371/journal.pbio.2003693 WOS:000426253300008. PMID: 29470495

27. Aarts AA, Anderson JE, Anderson CJ, Attridge PR, Attwood A, Axt J, et al. Estimating the reproducibility of psychological science. Science. 2015; 349(6251). https://doi.org/10.1126/science.aac4716 WOS:000360646800042. PMID: 26315443

28. Camerer CF, Dreber A, Forsell E, Ho TH, Huber J, Johannesson M, et al. Evaluating replicability of laboratory experiments in economics. Science. 2016; 351(6280):1433–1436. https://doi.org/10.1126/science.aaf0918 WOS:000372756200044. PMID: 26940865

29. Camerer CF, Dreber A, Holzmeister F, Ho TH, Huber J, Johannesson M, et al. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. Nat Hum Behav. 2018; 2(9):637–644. https://doi.org/10.1038/s41562-018-0399-z WOS:000446615600020. PMID: 31346273

30. Errington TM, Mathur M, Soderberg CK, Denis A, Perfito N, Iorns E, et al. Investigating the replicability of preclinical cancer biology. Elife. 2021; 10. https://doi.org/10.7554/eLife.71601 WOS:000730075900001. PMID: 34874005

31. Freedman LP, Cockburn IM, Simcoe TS. The Economics of Reproducibility in Preclinical Research. PLoS Biol. 2015; 13(6). https://doi.org/10.1371/journal.pbio.1002165 WOS:000357339600005. PMID: 26057340

32. Kvarven A, Stromland E, Johannesson M. Comparing meta-analyses and preregistered multiple-laboratory replication projects. Nat Hum Behav. 2020; 4(4):423–434. https://doi.org/10.1038/s41562-019-0787-z WOS:000507725500002. PMID: 31873200

33. Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, et al. Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci. 2013; 14(5):365–376. https://doi.org/10.1038/nrn3475 WOS:000317913900012. PMID: 23571845

34. Lamberink HJ, Otte WM, Sinke MRT, Lakens D, Glasziou PP, Tijdink JK, et al. Statistical power of clinical trials increased while effect size remained stable: an empirical analysis of 136,212 clinical trials between 1975 and 2014. J Clin Epidemiol. 2018; 102:123–128. https://doi.org/10.1016/j.jclinepi.2018.06.014 WOS:000444662400014. PMID: 29981870

35. Ioannidis JPA, Stanley TD, Doucouliagos H. The Power of Bias in Economics Research. Econ J. 2017; 127(605):F236–F265. https://doi.org/10.1111/ecoj.12461 WOS:000418017100011.

36. Stanley TD, Carter EC, Doucouliagos H. What Meta-Analyses Reveal About the Replicability of Psychological Research. Psychol Bull. 2018; 144(12):1325–1346. https://doi.org/10.1037/bul0000169 WOS:000451031000006. PMID: 30321017

37. Yang Y, Hillebrand H, Lagisz M, Cleasby I, Nakagawa S. Low statistical power and overestimated anthropogenic impacts, exacerbated by publication bias, dominate field studies in global change biology. Glob Chang Biol. 2022; 28(3):969–989. Epub 2021/11/05. https://doi.org/10.1111/gcb.15972 PMID: 34736291.

38. Yang YF, Sánchez-Tójar A, O'Dea RE, Noble DWA, Koricheva J, Jennions MD, et al. Publication bias impacts on effect size, statistical power, and magnitude (Type M) and sign (Type S) errors in ecology and evolutionary biology. BMC Biol. 2023; 21(1). https://doi.org/10.1186/s12915-022-01485-y WOS:000964320800004. PMID: 37013585

39. Kimmel K, Avolio ML, Ferraro PJ. Empirical evidence of widespread exaggeration bias and selective reporting in ecology. Nat Ecol Evol. 2023. https://doi.org/10.1038/s41559-023-02144-3 WOS:001042001400003. PMID: 37537387

40. Serdar CC, Cihan M, Yucel D, Serdar MA. Sample size, power and effect size revisited: simplified and practical approaches in pre-clinical, clinical and laboratory studies. Biochem Medica. 2021; 31(1). https://doi.org/10.11613/Bm.2021.010502 WOS:000659914800001. PMID: 33380887

41. Wilson BM, Harris CR, Wixted JT. Science is not a signal detection problem. Proc Natl Acad Sci. 2020; 117(11):5559–5567. https://doi.org/10.1073/pnas.1914237117 PMID: 32127477

42. Forstmeier W, Schielzeth H. Cryptic multiple hypotheses testing in linear models: overestimated effect sizes and the winner's curse. Behav Ecol Sociobiol. 2011; 65(1):47–55. https://doi.org/10.1007/s00265-010-1038-5 WOS:000285786000005. PMID: 21297852

43. Palmer C, Pe'er I. Statistical correction of the Winner's Curse explains replication variability in quantitative trait genome-wide association studies. PLoS Genet. 2017; 13(7). https://doi.org/10.1371/journal.pgen.1006916 WOS:000406615300049. PMID: 28715421

44. Monsarrat P, Vergnes JN. The intriguing evolution of effect sizes in biomedical research over time: smaller but more often statistically significant. Gigascience. 2017; 7(1). https://doi.org/10.1093/gigascience/gix121 WOS:000425086500001. PMID: 29228281

45. Goldacre B. Bad pharma: how drug companies mislead doctors and harm patients. London: Fourth Estate; 2012.

46. Flint J, Munafo MR. Candidate and non-candidate genes in behavior genetics. Curr Opin Neurobiol. 2013; 23(1):57–61. https://doi.org/10.1016/j.conb.2012.07.005 WOS:000314562900010. PMID: 22878161

47. Phillips B, Haschler TN, Karp NA. Statistical simulations show that scientists need not increase overall sample size by default when including both sexes in in vivo studies. PLoS Biol. 2023; 21(6):e3002129. https://doi.org/10.1371/journal.pbio.3002129 PMID: 37289836

48. Gelman A, Hill J, Vehtari A. Regression and other stories: Cambridge University Press; 2020.

49. Siviter H, Bailes EJ, Martin CD, Oliver TR, Koricheva J, Leadbeater E, et al. Agrochemicals interact synergistically to increase bee mortality. Nature. 2021; 596(7872). https://doi.org/10.1038/s41586-021-03787-7 WOS:000681278500004. PMID: 34349259

50. Lakens D. Performing high-powered studies efficiently with sequential analyses. Eur J Soc Psychol. 2014; 44(7):701–710. https://doi.org/10.1002/ejsp.2023 WOS:000346557400006.

51. Lakens D. Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses. Soc Psychol Pers Sci. 2017; 8(4):355–362. https://doi.org/10.1177/1948550617697177 WOS:000405075500001. PMID: 28736600

52. Jones SR, Carley S, Harrison M. An introduction to power and sample size estimation. Emerg Med J. 2003; 20(5):453–458. https://doi.org/10.1136/emj.20.5.453 WOS:000185103400021. PMID: 12954688

53. Bissonette JA. Small sample size problems in wildlife ecology: a contingent analytical approach. Wildlife Biol. 1999; 5(2):65–71. WOS:000080973800001.

54. Voelkl B, Altman NS, Forsman A, Forstmeier W, Gurevitch J, Jaric I, et al. Reproducibility of animal research in light of biological variation. Nat Rev Neurosci. 2020; 21(7):384–393. https://doi.org/10.1038/s41583-020-0313-3 WOS:000537344600001. PMID: 32488205

55. Usui T, Macleod MR, McCann SK, Senior AM, Nakagawa S. Meta-analysis of variation suggests that embracing variability improves both replicability and generalizability in preclinical research. PLoS Biol. 2021; 19(5). https://doi.org/10.1371/journal.pbio.3001009 WOS:000664237300004. PMID: 34010281

56. Voelkl B, Wurbel H. Reproducibility Crisis: Are We Ignoring Reaction Norms? Trends Pharmacol Sci. 2016; 37(7):509–510. https://doi.org/10.1016/j.tips.2016.05.003 WOS:000378961500001. PMID: 27211784

57. Debat V, David P. Mapping phenotypes: canalization, plasticity and developmental stability. Trends Ecol Evol. 2001; 16(10):555–561. https://doi.org/10.1016/S0169-5347(01)02266-2 WOS:000171174800011.

58. Karp NA. Reproducible preclinical research-Is embracing variability the answer? PLoS Biol. 2018; 16 (3). https://doi.org/10.1371/journal.pbio.2005413 WOS:000428987600022. PMID: 29505576

59. van der Staay FJ, Arndt SS, Nordquist RE. The standardization-generalization dilemma: a way out. Genes Brain Behav. 2010; 9(8):849–855. https://doi.org/10.1111/j.1601-183X.2010.00628.x WOS:000283726100001. PMID: 20662940

60. Wurbel H, Voelkl B, Altman NS, Forsman A, Forstmeier W, Gurevitch J, et al. Reply to 'It is time for an empirically informed paradigm shift in animal research'. Nat Rev Neurosci. 2020; 21(11):661–662. https://doi.org/10.1038/s41583-020-0370-7 WOS:000561521200001. PMID: 32826978

61. Richter SH, Garner JP, Auer C, Kunert J, Wurbel H. Systematic variation improves reproducibility of animal experiments. Nat Methods. 2010; 7(3):167–168. https://doi.org/10.1038/nmeth0310-167 WOS:000275058200003. PMID: 20195246

62. Richter SH, Garner JP, Wurbel H. Environmental standardization: cure or cause of poor reproducibility in animal experiments? Nat Methods. 2009; 6(4):257–261. https://doi.org/10.1038/nmeth.1312 WOS:000264738800012. PMID: 19333241

63. Langdalen H, Abrahamsen EB, Selvik JT. On the importance of systems thinking when using the ALARP principle for risk management. Reliab Eng Syst Safe. 2020; 204. https://doi.org/10.1016/j.ress.2020.107222 WOS:000583913400069.

64. Gelman A. Statistical Modeling, Causal Inference, and Social Science (https://statmodelingstatcolumbiaedu/2023/06/22/here-are-some-ways-of-making-your-study-replicable-no-its-not-what-you-think/) [Internet]. Available from: https://statmodeling.stat.columbia.edu/2023/06/22/here-are-some-ways-of-making-your-study-replicable-no-its-not-what-you-think/2023. [cited 2023].

65. Ryan TP. Modern experimental design. Hoboken, N.J.: Wiley-Interscience; Chichester: John Wiley [distributor]; 2007.

66. Herzog M. Understanding statistics and experimental design: how to not lie with statistics. New York, NY: Springer Berlin Heidelberg; 2019. pages cm p.

67. Lazic SE, Mellor JR, Ashby MC, Munafo MR. A Bayesian predictive approach for dealing with pseudoreplication. Sci Rep-Uk. 2020; 10(1). https://doi.org/10.1038/s41598-020-59384-7 WOS:000562858200017. PMID: 32047274

68. Colegrave N, Ruxton GD. Using Biological Insight and Pragmatism When Thinking about Pseudoreplication. Trends Ecol Evol. 2018; 33(1):28–35. https://doi.org/10.1016/j.tree.2017.10.007 WOS:000419242100004. PMID: 29122382

69. Forstmeier W, Wagenmakers EJ, Parker TH. Detecting and avoiding likely false-positive findings—a practical guide. Biol Rev. 2017; 92(4):1941–1968. https://doi.org/10.1111/brv.12315 WOS:000412314400005. PMID: 27879038

70. Arnqvist G. Mixed Models Offer No Freedom from Degrees of Freedom. Trends Ecol Evol. 2020; 35 (4):329–335. https://doi.org/10.1016/j.tree.2019.12.004 PMID: 31982147

71. Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MHH, et al. Generalized linear mixed models: a practical guide for ecology and evolution. Trends Ecol Evol. 2009; 24(3):127–135. https://doi.org/10.1016/j.tree.2008.10.008 WOS:000264615200003. PMID: 19185386

72. Green P, MacLeod CJ. SIMR: an R package for power analysis of generalized linear mixed models by simulation. Methods Ecol Evol. 2016; 7(4):493–498. https://doi.org/10.1111/2041-210x.12504 WOS:000373950700012.

73. Johnson PCD, Barry SJE, Ferguson HM, Muller P. Power analysis for generalized linear mixed models in ecology and evolution. Methods Ecol Evol. 2015; 6(2):133–142. https://doi.org/10.1111/2041-210X.12306 WOS:000349628100002. PMID: 25893088

74. DeBruine LM, Barr DJ. Understanding Mixed-Effects Models Through Data Simulation. Adv Meth Pract Psych. 2021; 4(1). https://doi.org/10.1177/2515245920965119 WOS:000708952600001.

75. Bateson M, Martin PR. Measuring behaviour: an introductory guide. 4th ed. 2021.

76. Begg CB. In Defense of P Values. Jnci Cancer Spect. 2020; 4(2). https://doi.org/10.1093/jncics/pkaa012 WOS:000608017100009. PMID: 32373778

77. Murtaugh PA. In defense of P values. Ecology. 2014; 95(3):611–617. https://doi.org/10.1890/13-0590.1 WOS:000332823100005. PMID: 24804441

78. Rice DB, Moher D. Curtailing the Use of Preregistration: A Misused Term. Perspect Psychol Sci. 2019; 14(6):1105–1108. https://doi.org/10.1177/1745691619858427 WOS:000483904100001. PMID: 31449761

79. Schmucker C, Schell LK, Portalupi S, Oeller P, Cabrera L, Bassler D, et al. Extent of Non-Publication in Cohorts of Studies Approved by Research Ethics Committees or Included in Trial Registries. PLoS ONE. 2014; 9(12). https://doi.org/10.1371/journal.pone.0114023 WOS:000348563300010. PMID: 25536072

80. Allen C, Mehler DMA. Open science challenges, benefits and tips in early career and beyond. PLoS Biol. 2019; 17(5). https://doi.org/10.1371/journal.pbio.3000246 WOS:000470189800010. PMID: 31042704

81. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship (vol 15, 160018, 2016). Sci Data. 2019; 6. https://doi.org/10.1038/s41597-019-0009-6 WOS:000464192900001. PMID: 30890711

82. DeVito NJ, Bacon S, Goldacre B. Compliance with legal requirement to report clinical trial results on ClinicalTrials.gov: a cohort study. Lancet. 2020; 395(10221):361–369. https://doi.org/10.1016/S0140-6736(19)33220-9 WOS:000510860200036. PMID: 31958402

83. Goldacre B, DeVito NJ, Heneghan C, Irving F, Bacon S, Fleminger J, et al. Compliance with requirement to report results on the EU Clinical Trials Register: cohort study and web resource. Bmj-Brit Med J. 2018; 362. https://doi.org/10.1136/bmj.k3218 WOS:000445016500001. PMID: 30209058

84. Jeffers MS, Maclellan A, Avey MT, Menon JM, Sunohara-Neilson J, Fergusson DA, et al. A call to implement preclinical study registration in animal ethics review. PLoS Biol. 2023; 21(10). https://doi.org/10.1371/journal.pbio.3002293 WOS:001082506500002. PMID: 37796782

85. Collins SL. Pluralism in Ecological Research. Bioscience. 2022; 72(10):927. https://doi.org/10.1093/biosci/biac089

86. Davies SW, Putnam HM, Ainsworth T, Baum JK, Bove CB, Crosby SC, et al. Promoting inclusive metrics of success and impact to dismantle a discriminatory reward system in science. PLoS Biol. 2021; 19(6). WOS:000665479600001. https://doi.org/10.1371/journal.pbio.3001282 PMID: 34129646

87. Amaral OB, Neves K. Reproducibility: expect less of the scientific paper Comment. Nature. 2021; 597(7876):329–331. https://doi.org/10.1038/d41586-021-02486-7 WOS:000696334600004. PMID: 34526702

88. Ioannidisa JPA. Meta-research: The art of getting it wrong. Res Synth Methods. 2010; 1(3–4):169–184. https://doi.org/10.1002/jrsm.19 WOS:000209380500001. PMID: 26061464

89. Trisos CH, Auerbach J, Katti M. Decoloniality and anti-oppressive practices for a more ethical ecology. Nat Ecol Evol. 2021; 5(9):1205–1212. WOS:000653692400001. https://doi.org/10.1038/s41559-021-01460-w PMID: 34031567

90. Nakagawa S, Lagisz M. Next steps after airing disagreement on a scientific issue with policy implications: a meta-analysis, multi-lab replication and adversarial collaboration. BMC Biol. 2023; 21(1). https://doi.org/10.1186/s12915-023-01567-5 WOS:000992782300001. PMID: 37217976

91. Seidler AL, Hunter KE, Cheyne S, Ghersi D, Berlin JA, Askie L. A guide to prospective meta-analysis. Bmj-Brit Med J. 2019; 367. https://doi.org/10.1136/bmj.l5342 WOS:000490448800001. PMID: 31597627

92. Coles NA, Hamlin JK, Sullivan LL, Parker TH, Altschul D. Build up big-team science. Nature. 2022; 601(7894):505–507. Epub 2022/01/27. https://doi.org/10.1038/d41586-022-00150-2 PMID: 35079150.

93. Frank MC, Alcock KJ, Arias-Trejo N, Aschersleben G, Baldwin D, Barbu S, et al. Quantifying Sources of Variability in Infancy Research Using the Infant-Directed-Speech Preference. Adv Meth Pract Psych. 2020; 3(1):24–52. https://doi.org/10.1177/2515245919900809 WOS:000710531200002.

94. Culina A, Adriaensen F, Bailey LD, Burgess MD, Charmantier A, Cole EF, et al. Connecting the data landscape of long-term ecological studies: The SPI-Birds data hub. J Anim Ecol. 2021; 90(9):2147–2160. https://doi.org/10.1111/1365-2656.13388 WOS:000595924500001. PMID: 33205462

95. Borer ET, Grace JB, Harpole WS, MacDougall AS, Seabloom EW. A decade of insights into grassland ecosystem responses to global environmental change. Nat Ecol Evol. 2017; 1(5). https://doi.org/10.1038/s41559-017-0118 WOS:000417173100008. PMID: 28812706

96. Wolf S, Mahecha MD, Sabatini FM, Wirth C, Bruelheide H, Kattge J, et al. Citizen science plant observations encode global trait patterns. Nat Ecol Evol. 2022; 6(12):1850–+. https://doi.org/10.1038/s41559-022-01904-x WOS:000870685300004. PMID: 36266458

97. Ebersole CR, Atherton OE, Belanger AL, Skulborstad HM, Allen JM, Banks JB, et al. Many Labs 3: Evaluating participant pool quality across the academic semester via replication. J Exp Soc Psychol. 2016; 67:68–82. https://doi.org/10.1016/j.jesp.2015.10.012 WOS:000384398700012.

98. Klein RA, Ratliff KA, Vianello M, Adams RB, Bahnik S, Bernstein MJ, et al. Investigating Variation in Replicability A "Many Labs" Replication Project. Soc Psychol-Germany. 2014; 45(3):142–152. https://doi.org/10.1027/1864-9335/a000178 WOS:000336836900002.

99. Stanley TD, Doucouliagos H. Meta-regression approximations to reduce publication selection bias. Res Synth Methods. 2014; 5(1):60–78. https://doi.org/10.1002/jrsm.1095 WOS:000348585200005. PMID: 26054026

100. Elliott JH, Synnot A, Turner T, Simmonds M, Akl EA, McDonald S, et al. Living systematic review: 1. Introduction-the why, what, when, and how. J Clin Epidemiol. 2017; 91:23–30. https://doi.org/10.1016/j.jclinepi.2017.08.010 WOS:000417550400005. PMID: 28912002

101. Nakagawa S, Dunn AG, Lagisz M, Bannach-Brown A, Grames EM, Sanchez-Tojar A, et al. A new ecosystem for evidence synthesis. Nat Ecol Evol. 2020; 4(4):498–501. WOS:000521526800005. https://doi.org/10.1038/s41559-020-1153-2 PMID: 32203483

102. Nakagawa S, Ivimey-Cook ER, Grainger MJ, O'Dea RE, Burke S, Drobniak SM, et al. Method Reporting with Initials for Transparency (MeRIT) promotes more granularity and accountability for author contributions. Nat Commun. 2023; 14(1). https://doi.org/10.1038/s41467-023-37039-1 WOS:001002031500015. PMID: 37012240

103. Coles NA, DeBruine LM, Azevedo F, Baumgartner HA, Frank MC. 'Big team' science challenges us to reconsider authorship. Nat Hum Behav. 2023; 7(5):665–667. https://doi.org/10.1038/s41562-023-01572-2 WOS:000952874800007. PMID: 36928785

104. Munafò MR, Chambers C, Collins A, Fortunato L, Macleod M. The reproducibility debate is an opportunity, not a crisis. BMC Res Notes. 2022; 15(1):43. https://doi.org/10.1186/s13104-022-05942-3 PMID: 35144667