Statistical Reports

Ecology, 97(10), 2016, pp. 2554–2561
© 2016 by the Ecological Society of America

Underappreciated problems of low replication in ecological field studies

Nathan P. Lemoine,[1] Ava Hoffman, Andrew J. Felton, Lauren Baur, Francis Chaves, Jesse Gray, Qiang Yu,[2] and Melinda D. Smith

Department of Biology, Graduate Degree Program in Ecology, Colorado State University, Fort Collins, Colorado 80523 USA

Abstract. The cost and difficulty of manipulative field studies makes low statistical power a pervasive issue throughout most ecological subdisciplines. Ecologists are already aware that small sample sizes increase the probability of committing Type II errors. In this article, we address a relatively unknown problem with low power: underpowered studies must overestimate small effect sizes in order to achieve statistical significance. First, we describe how low replication coupled with weak effect sizes leads to Type M errors, or exaggerated effect sizes. We then conduct a meta-analysis to determine the average statistical power and Type M error rate for manipulative field experiments that address important questions related to global change; global warming, biodiversity loss, and drought. Finally, we provide recommendations for avoiding Type M errors and constraining estimates of effect size from underpowered studies.

Key words: Bayesian statistics; LASSO regression; power; priors; ridge regression; Type M error; Type S error.

Introduction

The pervasiveness of global changes requires that ecologists accurately quantify the consequences of global change on community structure and ecosystem function. To obtain such estimates, ecologists simulate one or multiple aspects of global change (e.g., warming, biodiversity loss, or drought) using manipulative field experiments. Yet field experiments often yield equivocal results. Biodiversity has been reported as having variable effects on ecosystem functioning (Cardinale et al. 2011, Wardle 2016). Likewise, warming can stimulate (Cantarel et al. 2013), reduce (Cantarel et al. 2013), or have no effect (Biasi et al. 2008) on aboveground net primary productivity (ANPP), whereas drought effect sizes vary from small to large (Beier et al. 2012). Undoubtedly, inconsistent results among studies arise partly from methodological differences, differing biotic/abiotic contexts, and temporal variation. However, we contend that contradictory results may also arise even in perfectly replicated studies as a consequence of low statistical power that by necessity plagues many global change experiments.

Importantly, we do not focus on the canonical definition of low power as the failure to correctly reject the null hypothesis (i.e., Type II error). That issue has been addressed repeatedly in ecological sciences (Peterman 1990, Taylor and Gerrodette 1993, Jennions and Møller 2003, Nakagawa 2004). Instead, we focus on the recent realization that low-powered experiments examining processes that have small true effect sizes (μ_{true}) have a large probability of obtaining of a result of the wrong sign (i.e., a negative estimate of a positive μ_{true}, Type S error) and, by extension, must observe an overestimated effect size (μ_{obs}) in order to achieve statistical significance (i.e., Type M error; Ioannidis 2005, Button et al. 2013). The initial overestimate of effect sizes is referred to as the "winner's curse" because subsequent experiments using similar underpowered studies often cannot replicate the result because the statistical significance of the initial research occurred by chance (Young et al. 2008, Forstmeier and Schielzeth 2011, Button et al. 2013). The lack of reproducibility can intensify the debate over the ecological consequences of global change, making it imperative to raise awareness of this lesser known aspect of low power. Here, we describe this overlooked aspect of power, examine its prevalence in global change studies, and provide ways to remedy this important issue in ecological studies.

Manuscript received 12 April 2016; revised 7 June 2016; accepted 23 June 2016. Corresponding Editor: B. D. Inouye.

[2] Present address: Institute of Agricultural Resources and Regional Planning, Chinese Academy of Agricultural Sciences, Beijing 100081 China

[1] E-mail: lemoine.nathan@gmail.com

Table 1. Definitions of common terms.

Term	Definition
Type I error rate	the probability of incorrectly rejecting a true null hypothesis
Type II error rate	the probability of incorrectly accepting a false null hypothesis
Statistical power	the probability that a study correctly rejects the null hypothesis (1 − Type II error)
Effect size	the standardized change in a response variable (i.e., Cohen's *d*)
Critical value	the value of a test statistic (e.g., *t* value, chi-squared value) needed to achieve statistical significance at $\alpha = 0.05$
Type M error/Exaggeration ratio	the ratio of the critical value to the true effect size, such that Type M errors denote the magnitude by which an effect size must be overestimated in order to achieve significance
Winner's curse	the first, often high-profile paper, reports results that cannot be reproduced in subsequent experiments because the true effect size is small and the first study overestimated the effect size by chance

The Underappreciated Problem of Low Power

Most ecologists learn the well-established definition of statistical power in introductory statistics (Table 1). The Type II error rate (β) of an experiment is the probability that the analysis fails to reject the null hypothesis, H_0, when an effect is truly present (Fig. 1A). The probability of correctly rejecting H_0 $(1 - \beta)$ constitutes the statistical power of a study (Fig. 1A). Power increases with sample size (N) because sampling uncertainty of the effect size (σ_s) decreases with higher levels of replication ($\sigma_s = \sigma/\sqrt{N}$). However, there is a second yet often unrecognized consequence of statistical power: if μ_{true} is small, underpowered studies might reverse the sign of an effect and must considerably overestimate the magnitude of that effect to achieve significance at $P \leq 0.05$ (Young et al. 2008, Button et al. 2013).

Overestimates arise because the critical value (Z) for a given test is substantially larger than μ_{true} when μ_{true} is small and σ_s is large (Fig. 1B). Since $\mu_{\text{obs}} \geq Z$ is necessary for statistical significance, low power forces μ_{obs} to be much greater than μ_{true} (Fig. 1B). The ratio $Z{:}\mu_{\text{true}}$ is the Type M error rate and quantifies the proportion by which the critical value *must* exceed the effect size in order to achieve statistical significance. (Gelman and Carlin 2014). For example, consider an experiment with a small effect size but highly variable data: $\mu_{\text{true}} = 0.75$ and $\sigma = 4$. A study with $N = 10$ yields $\sigma_s = 1.27$ and $Z = 2.86$. Therefore, μ_{obs} must be ≥2.86 to achieve significance at $P \leq 0.05$, a Type M error of 3.8 (Fig. 1B). Additionally, highly uncertain σ_s contains both positive and negative tails, and the probability that μ_{obs} falls in the tail of incorrect sign is termed a Type S error. In this example, 28.4% of σ_s is negative despite a positive μ_{true}, such that the Type S error rate is 0.284 (Fig. 1B). Replicate studies with similar sample sizes will likely not find a significant result because most of the sampling distribution falls below the critical value (Fig. 1B), leading to the "winner's curse" and contradictory patterns in the literature.

Increasing replication to $N = 50$ significantly shrinks σ_s and Z, leading to higher power, an increased likelihood of repeatability, a lower Type S error rate of 0.096, and a lower Type M error of ~1.5 (Fig. 1C). Indeed, Type M errors rapidly decline to an asymptote as N increases

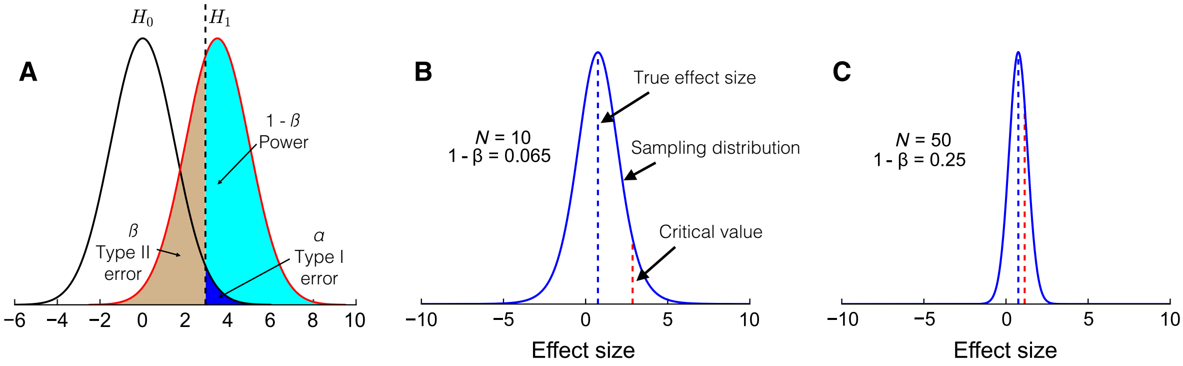

Fig. 1. (A) The Type I error rate (α) is the area of the null distribution (H_0) that falls beyond the critical value (dashed line). The Type II error rate (β) is the portion of the observed sampling distribution (H_1) that falls below the critical value. Statistical power is the portion of the observed sampling distribution that falls above the critical value. (B) If a study has low power, the critical value needed to achieve significance at $P \leq 0.05$ can be much larger than the true effect size, leading to overestimates of the effect size and irreproducible results. This example depicts $\mu_{\text{true}} = 0.75$ and standard deviation of the data $\sigma = 4$. With $N = 10$ replicates, the critical value is 2.86, which is over 3.8× larger than the true effect size. In this case, statistical power is 0.065. (C) Increasing replication to $N = 50$ yields increased power of 0.25. The critical value shrinks to 1.14, which is only ~1.5× larger than the true effect size. Panels B and C are adapted from http://andrewgelman.com/2014/11/17/power-06-looks-like-get-used/.

(Fig. 2), meaning that Type M errors decline rapidly with increasing power (Gelman and Carlin 2014). As a consequence, ecologists can minimize Type M errors with relatively low N, although the minimum Type M error for a study might still be quite large (Fig. 2). Naturally, large μ_{true}, low σ, and high N will all increase the power of a study and reduce potential Type M errors. The practical consequence is that ecologists must be cognizant of μ_{true}, or at least an expectation of its magnitude, prior to any statistical analyses. At a minimum, prior knowledge of μ_{true} allows ecologists to calculate Type M errors and raises awareness of potential overestimates. At best, ecologists can incorporate prior estimates of μ_{true} into Bayesian analyses that yield a posterior distribution the balances prior and new information.

Identifying Low Power in Ecological Studies

To determine the extent to which Type M errors pervade ecological field studies, we conducted a meta-analysis of three important ecological questions related to the effects of global change on ecosystem function: (1) What are the effects of warming on plant growth or aboveground net primary production (ANPP)? (2) What are the effects of biodiversity on ANPP or soil resource concentrations? (3) How does drought affect plant biomass or ANPP? These span a range of potential ecosystem effects that we predicted would have small, medium, and large μ_{true}, respectively.

As a first step, we identified contemporary meta-analyses on each subject in order to obtain the relevant primary literature. We then supplemented each meta-analysis with a second literature search using Web of Science in order to update each database to November 2015 (Appendix S1: Fig. S1, Appendix S3: Data S1). We conducted a separate meta-analysis for each question, which provided estimates of μ_{true}, σ_s, statistical power, and Type M error rate for each study (Button et al. 2013; Appendix S2, Appendix S3: Data S2).

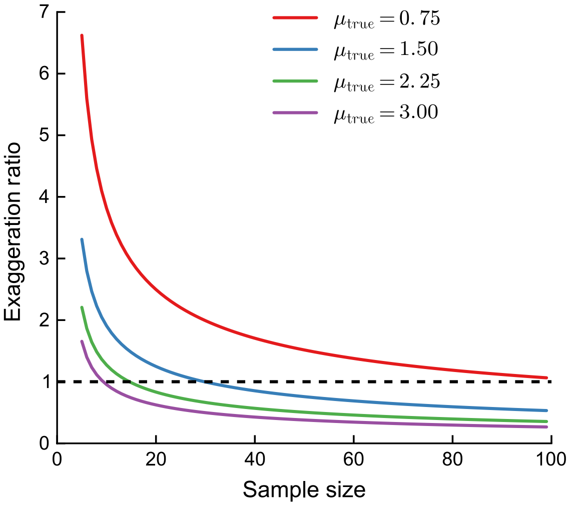

Fig. 2. Exaggeration ratio of the critical value needed for significance at various μ_{true}. Increasing sample size reduces the exaggeration ratio with diminishing returns. An exaggeration ratio of 1 indicates that the critical value equals the effect size (dashed line). All calculations were performed with $\sigma = 4$.

Warming Effects on ANPP

Overall, warming had a significant effect on plant biomass and growth ($P < 0.001$), although μ_{true} was small (0.56, CI_{95} = 0.42–0.69). Warming experiments had extraordinarily low statistical power due to the combination of low replication (N = 9 ± 7 per treatment), a small μ_{true}, and highly variable data (σ = 84.9 ± 93.1; Fig. 3). Average power was 0.06 ± 0.02, and the highest power was only 0.14. Interestingly, power had no relationship with the number of replicates ($P = 0.846$) but did have a positive correlation with duration ($P < 0.001$). However, the positive relationship between study duration and power was driven by three experiments from a single 22-yr study. Removal of this study negated any influence of duration on statistical power ($P = 0.488$). Type M errors were generally large, averaging 3.29 ± 0.23. As a consequence, studies examining warming effects on plant biomass must either commit a Type II error or overestimate μ_{true} by more than threefold in order to achieve statistical significance at $P \leq 0.05$.

Biodiversity–Ecosystem Function

Biodiversity effects on ecosystem productivity were, as expected, moderate ($P < 0.001$, μ_{true} = 1.22, CI_{95} = 0.96–1.48), yielding higher average power than warming experiments (0.22 ± 0.30). Seven studies achieved power >0.8, although 70% of experiments (40) had power < 0.2 (Fig. 3). This bimodality was driven mostly by three experiments with abnormally high power: the Cedar Creek LTER (BioCON, and Biodiversity II experiments) and BIODEPTH experiments, all of which had a large number of replicates. However, neither sample size ($P = 0.949$) nor duration ($P = 0.926$) were related to statistical power. Despite low power, the moderate μ_{true} of biodiversity yielded lower Type M error rates, averaging 1.42 ± 0.08. Such low Type M error rates suggest that biodiversity studies need only exaggerate the effect by ~1.5× in order to achieve statistical significance.

Drought Effects on ANPP

Overall, drought strongly affected ANPP ($P < 0.001$, μ_{true} = 2.91, CI_{95} = 1.79–4.03). Still, 61% of studies had power <0.1 and only 12% had power >0.5 (Fig. 3). Increased replication led to significantly higher power ($P < 0.001$), although this trend was driven by one study (two experiments) with 50 replicates. Removal of these negated any effect of replication on power ($P = 0.751$). Study duration had no effect on statistical power ($P = 0.829$). Despite low power, Type M error rates were

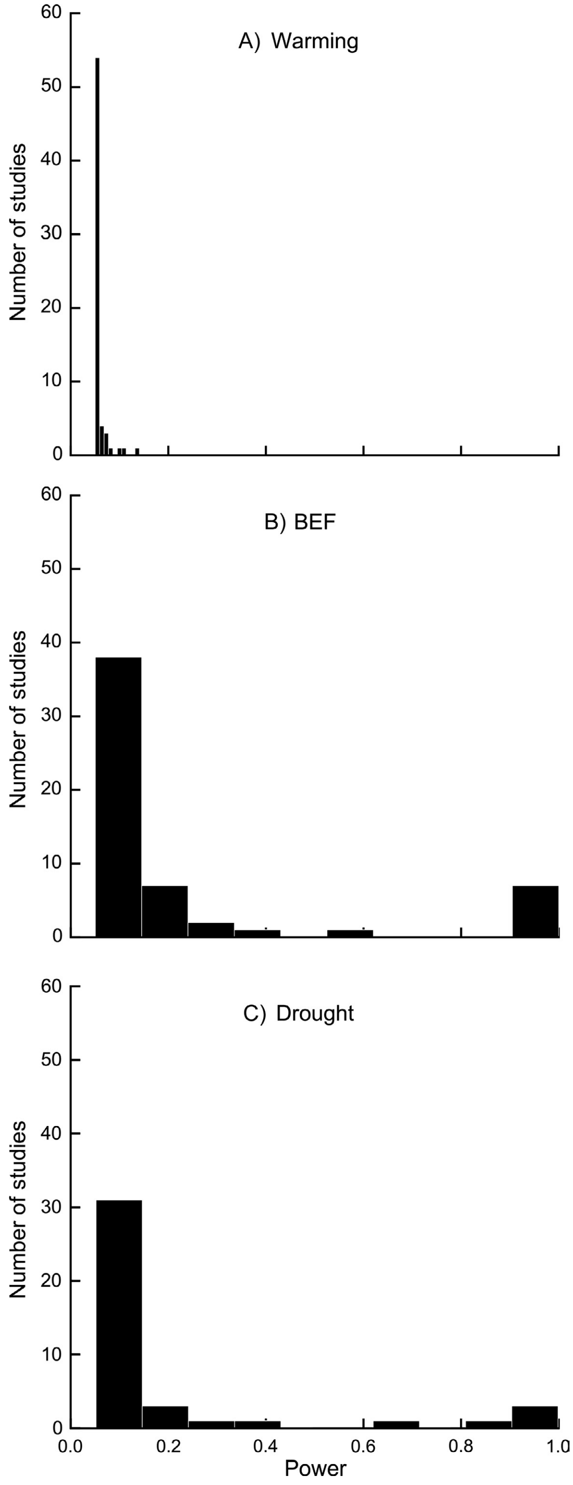

FIG. 3. Histograms of statistical power for warming, biodiversity–ecosystem function (BEF), and drought experiments.

typically low, averaging 0.66 ± 0.1. That Type M errors are <1 for most studies suggests that the overall effect size is nearly always greater than the critical value needed to detect statistical significance. As a result, an underpowered drought study is at greater risk of committing a Type II error than a Type M error.

OVERALL PATTERNS

Based on the above conclusions, it is worth considering what factors yield high power and low Type M errors. Interestingly, statistical power did not depend on either the difference between treatments or replication, but instead declined sharply as data variance increased (Fig. 4A). Type M error rates, however, did not depend upon data variability but instead declined within increasing sample size, as expected (Figs. 2, 4B). Also as expected, Type M error rates were lowest for those responses with the largest μ_{true} (Fig. 4B).

REMEDIES AND FUTURE DIRECTIONS

Ecological field experiments consistently demonstrated low statistical power. Such low power has long been identified as problematic with respect to Type II error rates (Peterman 1990, Jennions and Møller 2003, Wardle 2016). Here, we described a second, underappreciated aspect of low power: poorly replicated studies must overestimate μ_{true} in order to achieve statistical significance (Ioannidis 2005, Young et al. 2008, Button et al. 2013). The relevant question is then: what are the best practices to maximize statistical power and minimize Type M errors?

High sample sizes are ostensibly the most obvious answer. Indeed, increased replication reduced Type M errors as expected (Figs. 2, 4B), but sample size had no relationship with statistical power (Appendix S1: Figure S2). Instead, power was lowest for highly variable data, a common aspect of field experiments that is difficult for ecologists to mitigate. Furthermore, increasing *N* can only partially alleviated Type M errors in the face of high data variability and small μ_{true} (Figs. 2, 4B). Finally, levels of replication required to achieve adequate power can be cost-prohibitive, with $N > 100$ necessary to minimize Type M errors and maximize power in many cases (similar to values reported by Button et al. 2013). Given that increased sample sizes can only partially offset the issues of low power and Type M errors, we must seek changes to the underlying principles of reporting and judging ecological results.

Previous authors advocated that nonsignificant findings be complemented by "observed power" calculations (Peterman 1990, Jennions and Møller 2003). These calculations assume that μ_{obs} and σ_s^2 in a study represent μ_{true} and σ_s^2 for use in power calculations (e.g., Lemoine and Valentine 2012). We strongly advise against this circular approach because "observed power" is a direct function of the *P* value; low *P* values yield high power by default and vice versa (Hoenig and Heisey 2001). "Observed power" therefore provides no new information beyond the *P*-value. Instead, we propose three changes to reporting of results and statistical practice that help offset the overestimation of μ_{true}.

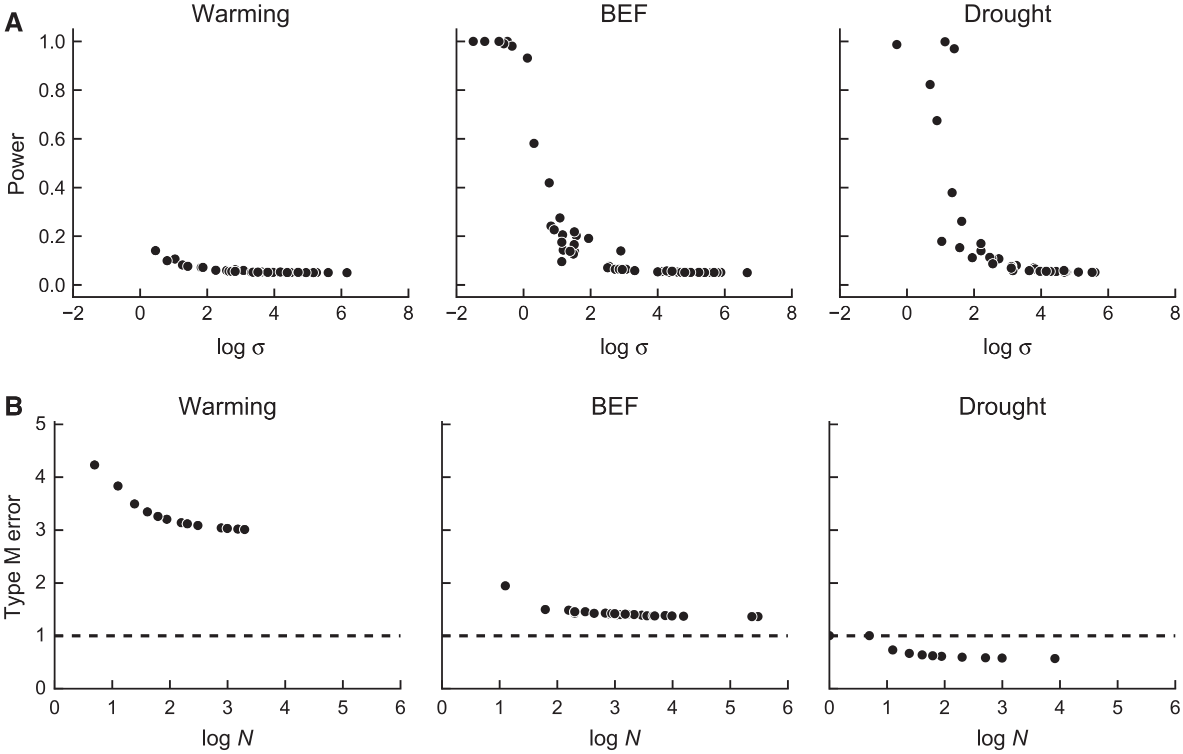

FIG. 4. (A) Relationship between statistical power and the pooled standard deviation σ of each study (see Appendix S2 and Appendix S3: Data S2 for calculations). (B) Relationship between Type M error and sample size N for each study. Dashed line indicates a Type M error, or exaggeration ratio, of 1, where the critical value equals the effect size.

Calculate type M error

Instead of "observed power," we recommend that ecologists calculate the Type M error. This quantity is independent of the reported P value and instead relies on an estimate of μ_{true} obtained from independent sources (e.g., meta-analysis). If no or few previous studies are available, researchers can use single studies or pilot studies to inform μ_{true}. Calculating the Type M error provides three advantages: First, by requiring ecologists to estimate μ_{true}, Type M error calculations force ecologists to be aware of the expected effect size and the statistical power of their own study. Second, emphasizing Type M errors reminds ecologists not to interpret μ_{obs} as immutable but rather to understand that μ_{obs} is subject to chance and uncertainty. Third, Type M error calculations provide an easily interpretable metric to determine if findings exaggerate the true effect. Scripts in the R statistical language are readily available for Type M error calculations and examples are given in Appendix S3: Data S3 (Gelman and Carlin 2014).

A statement of Type M error would ideally follow reporting of significant results in low-powered studies. For example, "Warming significantly increased the number of flowers produced by *Oenothera biennis* over the growing season by 200 ± 40 (P < 0.001). However, given our small sample size and that most effects of warming on flower production are considerably weaker, we calculated the Type M error of our study as 1.7. The average increase in *O. biennis* flower production due to warming is therefore probably smaller than reported here." This statement neither reduces the impact nor weakens the conclusions of the experiment; warming still has a positive effect on flower production. Rather, the statement gives a more informative account that the true effect may be lower than the reported effect, which would enable subsequent experiments to accurately validate the effects of warming on flowering.

Remove statistical dichotomies; report effect sizes and uncertainty for all results

Reporting Type M errors, as suggested above, shifts the emphasis of statistical analyses from achieving statistical significance towards providing accurate estimates of effect size. In fact, many statistical issues, including Type II errors, Type M errors, and publication bias, stem from the current emphasis on dichotomous significance; results are either significant at $P \leq 0.05$ or considered unimportant. This philosophy relegates biological significance to a secondary concern behind statistical significance (Nakagawa and Cuthill 2007). We join others in calling for a shift in the statistical paradigm away from thresholds for statistical significance (Nakagawa and Cuthill 2007, Hurlbert and Lombardi 2009, Cumming 2014) and argue that ecologists should emphasize effect sizes and confidence intervals (Colegrave and Ruxton 2003). Even R. A. Fisher eventually acknowledged that *P* values should be interpreted fluidly as a measure of relative support for the null hypothesis (Hurlbert and Lombardi 2009, Murtaugh 2014). However, we do not advocate entirely abandoning *P* values in favor of other methods, e.g., information criteria (IC), as others recently have (Anderson and Burnham 2002, Barber and Ogle 2014, Burnham and Anderson 2014). Researchers often subject IC and Bayesian methods to arbitrary thresholds of significance similar to the $P \leq 0.05$ rule of significance, i.e., $\Delta\text{AIC} < 2$ for IC methods (e.g., Lemoine et al. 2014) or Bayesian significance testing using 95% posterior credible intervals (Gelman and Hill 2007, Kruschke 2010). We propose that the algorithm used to estimate parameters (e.g., least squares, maximum likelihood, Markov chains) is irrelevant when the results are subject to significance testing (Forstmeier and Schielzeth 2011). Rather, the major philosophical advance is to avoid binning results into "significant" and "not significant" categories based on an arbitrary value of any summary statistic.

By tempering significance testing (i.e., $P \leq 0.05$) as the primary criterion for publication and scientific impact, ecologists can shift the focus of their analyses from statistical significance to effect sizes (Nakagawa and Cuthill 2007). Emphasizing accurate estimates of effect sizes would prioritize minimizing Type M as opposed to Type II errors. This philosophy inherently forces ecologists to discuss the magnitude and "biological significance" of their results as opposed to categorical differences, and will encourage placing results in the context of previous effect sizes (Nakagawa and Cuthill 2007). The practical implication would be for researchers to complement all *P* values with effect sizes and confidence intervals, including post hoc multiple comparisons (Colegrave and Ruxton 2003). All too often, nonsignificant results are dismissed with no mention of effect sizes. Yet, nonsignificant results could potentially be as important as results that meet the $P \leq 0.05$ criteria.

Consider an experiment designed to assess the effects of warming on flower production of *Oenothera biennis* with $n = 5$ per temperature treatment. Over the course of a growing season, *O. biennis* produces ~900 ± 100 flowers at ambient temperatures (Lemoine et al., *unpublished manuscript*). Suppose warming increases the number of flowers produced throughout the course of the growing season by 20 ± 6, ($\text{CI}_{95} = 3.3$–36.7). Although this result is over three standard deviations from zero and significant at $P = 0.002$, it represents a negligible fraction of total flower production (~2%). Suppose instead that warming increases flower production by 300 ± 275, a result that is not significant at $P \leq 0.05$ ($\text{CI}_{95} = -463$ to 1,063.5). Because $P = 0.07$, the CI_{95} includes zero but still indicates that it is highly possible that warming will substantially increase the number of flowers, potentially by 100%. In this example, the latter situation represents a potentially larger effect on *O. biennis* flowering than the former situation but would be dismissed as less important if adhering strictly to criteria based on $P \leq 0.05$. Thus, reporting effect sizes and confidence intervals can help ecologists assess the biological importance of both significant and non-significant results.

Bayesian statistics

Although potentially more controversial, Bayesian statistics can resolve many of the issues related to low statistical power and are becoming increasingly popular for testing a variety of ecological hypotheses (Ellison 2004, Arhonditsis et al. 2006, Price et al. 2009, Thomson et al. 2010, Vieilledent et al. 2010). One of the principal advantages of Bayesian statistics, and perhaps its most contentious issue, is the ease with which researchers can assign prior information to all parameters and effect sizes. However, the ability of priors to influence the results might make some ecologists understandably uneasy with Bayesian statistics.

To demonstrate the influence of priors on posterior distributions, it is first necessary to realize that the Bayesian posterior distribution of a parameter given the data, $\Pr(\theta|\,Y)$, is proportional to the product of the likelihood of the data given the parameter ($L(Y|\theta)$) and the prior density ($\Pr(\theta)$):

$$\Pr(\theta|\,Y) \propto L(Y|\theta)\Pr(\theta).$$

The posterior can be thought of as a weighted average of the likelihood and prior with weights corresponding to sample sizes. Consider a random variable y with an estimated mean $\bar{y}$ and a known standard deviation σ_{obs}. We are interested in estimating the posterior distribution of the mean of the data, μ_{post}. The posterior distribution for the mean is $N(\mu_{\text{post}}, \tau^2_{\text{post}})$, estimated by

$$\mu_{\text{post}} = \frac{\frac{1}{\sigma^2_{\text{prior}}}\mu_{\text{prior}} + \frac{N}{\sigma^2_{\text{obs}}}\bar{y}}{\frac{1}{\sigma^2_{\text{prior}}} + \frac{N}{\sigma^2_{\text{obs}}}}$$

and

$$\tau^2_{\text{post}} = \frac{1}{\frac{1}{\sigma^2_{\text{prior}}} + \frac{N}{\sigma^2_{\text{obs}}}}.$$

From these equations, it should be apparent that, as sample size (N) increases, the prior becomes less influential and the posterior distribution of the mean, μ_{post}, and sampling uncertainty, τ^2_{post}, are dominated by the data (Appendix S1: Fig. S3). In other words, the larger the sample size, the more strongly we update our prior information about parameter values. Although this is a simplistic example, the general principle holds for all Bayesian models and posterior distributions (Appendix S1: Fig. S4).

Currently, most practitioners use uninformative prior distributions that are flat over the relevant range of parameters (e.g., $N(0, 1{,}000)$ for unbounded continuous variables, Beta(1, 1) for probabilities, *U*(lower, upper) in the rare case that parameters have known lower and upper limits). With such priors, Bayesian posterior distributions converge to maximum likelihood estimates

Box 1. Example of Bayesian analyses using various priors.

Ideally, researchers would conduct multiple analyses using different priors to judge the sensitivity of results to the choice of priors. Indeed, priors can be thought of as components of the statistical model, to be judged and updated if the posteriors do not provide sensible results (Gelman and Shalizi 2013). Posterior sensitivity to priors is related to sample size; analyses based on smaller samples will be more sensitive to prior choice than larger samples. Here we use two real data sets to demonstrate how the prior choice influences the posteriors predictions of effect sizes for studies that differ in their level of replication.

Ewel et al. (2015) reported the biomass of tropical trees grown in mixtures and monocultures with three replicates. Their study reports an effect size of 5.4 and a sampling variance of 9.4, with an observed distribution of $N(5.4, 3.07^2)$. We can calculate posterior predictions of the effect size based on the equation desrcibed in the 'Bayesian statistics' section, assuming the sampling variance is fixed, using three sets of priors:

1. Using completely uninformative priors $N(0, 1{,}000^2)$, the estimated effect size, 5.4 ± 1.77, is statistically significant at $P = 0.002$. However, low power (0.10) and high Type M error (2.0) suggest that this effect size is likely an overestimate. Attempts to replicate these findings will likely fail.
2. Weakly uninformative priors of $N(0, 1^2)$ help constrain the effect size and prevent over estimates, as the effect size has shrunk to 1.31 ± 0.76, and is no longer significant ($P = 0.09$). Although no longer significant, the effect size is still informative and the bulk of the posterior lay above 0 ($\Pr(>0) = 0.96$) and 66% of the posterior falls between 0.5 and 2.0. Thus, the effect is likely important, even if not significant at $P \leq 0.05$.
3. Under the assumption of strongly informative priors of $N(1.21, 0.14^2)$ obtained from our meta-analysis, the effect size is 1.24 ± 0.14, which is significant at $P < 0.001$. Because the study has only three replicates and we possess strong prior information about the effect size, the posterior distribution largely reflects the prior and has not been strongly updated.

In contrast, Cook-Patton et al. (2011) described the effect of biodiversity on aboveground plant biomass, but had a larger number of replicates ($N = 60$). They reported an effect size of $N(0.28, 0.04^2)$. With uninformative priors, the effect size 0.28 ± 0.03 is significant at $P < 0.0001$, and these estimates do not differ under weakly informative priors $N(0, 1^2)$. With strongly informative priors ($N[1.21, 0.14^2]$), the effect size has increased somewhat to 0.31 ± 0.03, but the estimate is still dominated by the data.

(Gelman et al. 2013) and do not alleviate the issue of Type M error. Informative priors, however, constrain estimated effect sizes to believable values in the presence of small sample sizes and, when combined with thorough reporting of effect sizes described above, represent an ideal solution to the problem of Type M errors and the "winner's curse" (Hoenig and Heisey 2001).

The choice of priors is one of the most contentious issues in Bayesian statistics. Ideally, informative priors would be based on information regarding μ_{true} derived from meta analyses, literature searches, or preliminary experiments (Garamszegi et al. 2009). Analyses using strongly informative priors should always be coupled with analyses using weak or uninformative priors to demonstrate the sensitivity of conclusions to prior information. In the absence of prior information, we advocate standardizing the response variable and placing a weakly informative prior of $N(0, 1)$ on all effects. This prior assumes that most effects will be within one standard deviation of the mean and very few effects will be larger than two standard deviations, although careful consideration of the scale of predictors is necessary as well. In fact, Bayesian regression with standard normal priors is analogous to ridge regression, which penalizes overly large coefficient estimates. The more severe Laplace prior, which concentrates most of the prior distribution near 0, can be used in lieu of the standard normal prior and is identical to LASSO regression. In these regression techniques (Bayesian, ridge, LASSO), studies with small sample sizes and extremely large estimated effects will have posteriors shrunk towards 0 (Appendix S1: Figs. S3, S4, Box 1). Such shrinkage makes Bayesian analyses with informative priors more conservative than frequentist analyses and helps prevent the erroneous estimation of large effect sizes in underpowered studies (e.g., Lemoine and Shantz 2016).

By adhering to these suggestions, ecologists can avoid the pitfalls of overstating results arising from underpowered and poorly replicated experiments, which given the logistical and fiscal constraints of many studies is a common occurrence in ecology. The "winner's curse" leads to irreproducible research and can generate debate about contrasting results that, instead, may simply reflect sampling uncertainty associated with weak and variable effect sizes. At the very least, providing confidence intervals and Type M errors will help ecologists and policy-makers assess the true effect size and its potential variation. At most, Bayesian statistics provide researchers with the ability to constrain posteriors to believable values unless strongly supported by the data and backed up by numerous observations (Box 1). Ecologists can and should debate the true effect size, which may differ among ecosystems, with different experimental methods, or among different study organisms. What is not debatable, however, is that underpowered studies addressing issues with small true effect sizes must overestimate the size of the effect in order to find statistical significance. Ecologists need to be aware of this issue in order to avoid the pitfalls of Type M errors and irreproducible research.

Acknowledgments

This work was supported by a USDA NIFA-AFRI Postdoctoral Fellowship (2016-67012-25169) to N. P. Lemoine and a grant from the National Natural Science Foundation of China (41320104002) to Q. Y. and M. D. S. We would like to thank Brian Inouye and two anonymous reviewers for their helpful comments. Dr. Andrew Gelman provided much of the inspiration for this manuscript.

Literature Cited

Anderson, D. R., and K. P. Burnham. 2002. Avoiding pitfalls when using information-theoretic methods. Journal of Wildlife Management 66:912–918.

Arhonditsis, G. B., C. A. Stow, L. J. Steinberg, M. A. Kenney, R. C. Lathrop, S. J. McBride, and K. H. Reckhow. 2006. Exploring ecological patterns with structural equation modeling and Bayesian analysis. Ecological Modelling 192: 385–409.

Barber, J. J., and K. Ogle. 2014. To *P* or not to *P*? Ecology 95:621–626.

Beier, C., et al. 2012. Precipitation manipulation experiments—challenges and recommendations for the future. Ecology Letters 15:899–911.

Biasi, C., H. Meyer, O. Rusalimova, R. Hämmerle, C. Kaiser, C. Baranyi, H. Daims, N. Lashchinsky, P. Barsukov, and A. Richter. 2008. Initial effects of experimental warming on carbon exchange rates, plant growth and microbial dynamics of a lichen-rich dwarf shrub tundra in Siberia. Plant and Soil 307:191–205.

Burnham, K. P., and D. R. Anderson. 2014. *P* values are only an index to evidence: 20th- vs. 21st-century statistical science. Ecology 95:627–630.

Button, K. S., J. P. A. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. J. Robinson, and M. R. Munafò. 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nature Reviews Neuroscience 14:365–376.

Cantarel, A. A. M., J. M. G. Bloor, and J.-F. Soussana. 2013. Four years of simulated climate change reduces aboveground productivity and alters functional diversity in a grassland ecosystem. Journal of Vegetation Science 24:113–126.

Cardinale, B. J., K. L. Matulich, D. U. Hooper, J. E. K. Byrnes, J. E. Duffy, L. Gamfeldt, P. Balvanera, M. I. O'Connor, and A. Gonzalez. 2011. The functional role of producer diversity in ecosystems. American Journal of Botany 98:572–592.

Colegrave, N., and G. D. Ruxton. 2003. Confidence intervals are a more useful complement to nonsignificant tests than are power calculations. Behavioral Ecology 14:446–450.

Cook-Patton, S. C., S. H. McArt, A. L. Parachnowitsch, J. S. Thaler, and A. A. Agrawal. 2011. A direct comparison of the consequences of genotypic and species diversity on communities and ecosystem function. Ecology 92:915–923.

Cumming, G. 2014. The new statistics: why and how. Psychological Science 25:7–29.

Ellison, A. M. 2004. Bayesian inference in ecology. Ecology Letters 7:509–520.

Ewel, J. J., G. Celis, and L. Schreeg. 2015. Steeply increasing growth differential between mixture and monocultures of tropical trees. Biotropica 47:162–171.

Forstmeier, W., and H. Schielzeth. 2011. Cryptic multiple hypotheses testing in linear models: overestimated effect sizes and the winner's curse. Behavioral Ecology and Sociobiology 65:47–55.

Garamszegi, L. Z., et al. 2009. Changing philosophies and tools for statistical inferences in behavioral ecology. Behavioral Ecology 20:1363–1375.

Gelman, A., and J. B. Carlin. 2014. Beyond power calculations: assessing Type S (Sign) and Type M (Magnitude) errors. Perspectives on Psychological Science 9:641–651.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. 2013. Bayesian data analysis. Third edition. Chapman and Hall, New York, New York, USA.

Gelman, A., and J. Hill. 2007. Data analysis using regression and multilevel/hierarchical models. Cambridge University Press, New York, New York, USA.

Gelman, A., and C. R. Shalizi. 2013. Philosophy and the practice of Bayesian statistics. British Journal of Mathematical and Statistical Psychology 66:8–38.

Hoenig, J. M., and D. M. Heisey. 2001. The abuse of power: the pervasive fallacy of power calculations for data analysis. American Statistician 55:1–6.

Hurlbert, S. H., and C. M. Lombardi. 2009. Final collapse of the Neyman-Pearson decision theoretic framework and rise of the neoFisherian. Annals Zoologici Fennici 46:311–349.

Ioannidis, J. P. A. 2005. Why most published research findings are false. PLoS Medicine 2:e124.

Jennions, M. D., and A. P. Møller. 2003. A survey of the statistical power of research in behavior ecology and animal behavior. Behavioral Ecology 14:438–445.

Kruschke, J. K. 2010. Doing Bayesian data analysis: a tutorial with R and BUGS. First edition. Academic Press, Burlington, Masschusetts, USA.

Lemoine, N. P., S. T. Giery, and D. E. Burkepile. 2014. Differing nutritional constraints of consumers across ecosystems. Oecologia 174:1367–1376.

Lemoine, N. P., and A. A. Shantz. 2016. Increased temperature causes protein limitation by reducing the efficiency of nitrogen digestion in the ectothermic herbivore *Spodoptera exigua*. Physiological Entomology 41:143–151.

Lemoine, N. P., and J. F. Valentine. 2012. Structurally complex habitats provided by *Acropora palmata* influence ecosystem processes on a reef in the Florida Keys National Marine Sanctuary. Coral Reefs 31:779–786.

Murtaugh, P. A. 2014. In defense of *P* values. Ecology 95:611–617.

Nakagawa, S. 2004. A farewell to Bonferroni: the problems of low statistical power and publication bias. Behavioral Ecology 15:1044–1045.

Nakagawa, S., and I. C. Cuthill. 2007. Effect size, confidence interval and statistical significance: a practical guide for biologists. Biological Reviews 82:591–605.

Peterman, R. M. 1990. The importance of reporting statistical power: the forest decline and acidic deposition example. Ecology 71:2024–2027.

Price, C. A., K. Ogle, E. P. White, and J. S. Weitz. 2009. Evaluating scale models in biology using hierarchical Bayesian approaches. Ecology Letters 12:641–651.

Taylor, B. L., and T. Gerrodette. 1993. The use of statistical power in conservation biology: the vaquita and northern spotted owl. Conservation Biology 7:489–500.

Thomson, J. R., W. J. Kimmerer, L. R. Brown, K. B. Newman, R. Mac Nally, W. A. Bennett, F. Feyrer, and E. Fleishman. 2010. Bayesian change point analysis of abundance trends for pelagic fishes in the upper San Francisco Estuary. Ecological Applications 20:1431–1448.

Vieilledent, G., B. Courbaud, G. Kunstler, J.-F. Dhôte, and J. S. Clark. 2010. Individual variability in tree allometry determines light resource allocation in forest ecosystems: a hierarchical Bayesian approach. Oecologia 163:759–773.

Wardle, D. A. 2016. Do experiments exploring plant diversity-ecosystem functioning relationships inform how biodiversity loss impacts natural ecosystems? Journal of Vegetation Science 27:646–653.

Young, N. S., J. P. A. Ioannidis, and O. Al-Ubaydli. 2008. Why current publication practices may distort science. PLoS Medicine 5:1418–1422.

Supporting Information

Additional supporting information may be found in the online version of this article at http://onlinelibrary.wiley.com/doi/10.1002/ecy.1506/suppinfo