Are experiment sample sizes adequate to detect biologically important interactions between multiple stressors?

Benjamin J. Burgess[1]

Michelle C. Jackson[2]

David J. Murrell[1]*

[1] **Centre for Biodiversity and Environment Research, Department of Genetics, Evolution and Environment, University College London, Gower Street, London, WC1E 6BT, United Kingdom**

[2] **Department of Zoology, University of Oxford, Oxford, OX1 3SZ, United Kingdom**

***corresponding author (d.murrell@ucl.ac.uk)**

Temporary Running Head: Sample sizes, and biological importance in stressor interactions

Key words: Ecosystem drivers, Replicates, Interactions, Experimental Design, Critical Effect Size, Minimum Biological Effect of Interest, Additive Null Model, Multiplicative Null Model.

Abstract

As most ecosystems are being challenged by multiple, co-occurring stressors, an important challenge is to understand and predict how stressors interact to affect biological responses. A popular approach is to design factorial experiments that measure biological responses to pairs of stressors and compare the observed response to a null model expectation. Unfortunately, we believe experiment sample sizes are inadequate to detect most non-null stressor interaction responses, greatly hindering progress. Determination of adequate sample size requires (i) knowledge of the detection ability of the inference method being used, and (ii) a consideration of the smallest biologically meaningful deviation from the null expectation. However, (i) has not been investigated and (ii) is yet to be discussed. Consequently, typical experimental sample sizes are small (3-4 replicates), except for some individual-level studies where replicates may exceed 100. Re-analysis of recently published data highlights how small sample sizes can only detect very large deviations from the null model, implying many important non-null stressor-pair interactions are being missed, and how sample size only improves detection ability slowly. Critically, researchers still need to define the smallest interaction of interest, i.e. the lower limit for a biologically important interaction, which is likely to be system specific, meaning that a general guide is unavailable. However, we suspect that most experiments may require 20 or more replicates per treatment. Sample sizes could potentially be increased by focussing on individual-level responses to multiple stressors, or by forming coordinated networks of researchers to repeat experiments in larger-scale studies. Our analyses relate to the additive and multiplicative null models and we encourage similar analyses on other null models and inference methods for stressor interactions. Without knowledge of the detection abilities of the statistical tools at hand, or definition of the smallest meaningful interaction, we will undoubtedly continue to miss important ecosystem stressor interactions.

Introduction

Most, if not all, ecosystems are being impacted by multiple co-occurring stressors (e.g., climate change, invasive species, pollution), which are predominately anthropogenic in origin (Halpern et al. 2015; Beauchesne et al. 2021). As such, there is an urgent need to document and predict how these stressors (threats, drivers) interact given that they are capable of affecting individuals through to entire ecosystems (Jackson et al. 2021; Simmons et al. 2021; Sokolova 2021). However, current knowledge of how stressors interact to affect various systems is limited (Hodgson & Halpern 2019; Lemm et al. 2021). To date, progress has been driven by individual studies that have contributed to larger-scale meta-analyses, but relatively few generalisations are possible (Côté et al. 2016; Orr et al. 2020). This is perhaps not surprising given the broad range of ecosystems, taxonomic groups, and biological responses that have been considered (e.g., Ban et al., 2014; Burgess et al., 2021; Lange et al., 2018), but another contributory factor that has not been examined is the issue of adequate sample sizes in multiple stressor experiments.

We contest that many potentially important stressor-pair interactions are being missed simply due to low replication number. In order to design effective multiple stressor experiments that have adequate sample sizes, researchers must consider the trifecta of: i) resource costs, ii) the smallest stressor-pair interaction that can be detected (statistical power), and iii) the minimum biological effect of interest (Figure 1). However, we believe only resource costs normally factor into experimental design since the detection limits of the statistical tools commonly used in stressor interactions have not been quantified, and there has been no discussion on what a biologically important stressor interaction is. We define the smallest interaction of interest as the smallest biologically relevant deviation from the null expectation and could represent the smallest deviation that would warrant a change in management strategy compared to the null. Here we

will look at sample sizes typical of stressor interaction experiments, use empirical examples and analysis of statistical models to highlight why it is likely important interactions are being missed, and show how the minimum biological effect of interest dictates the sample sizes required.

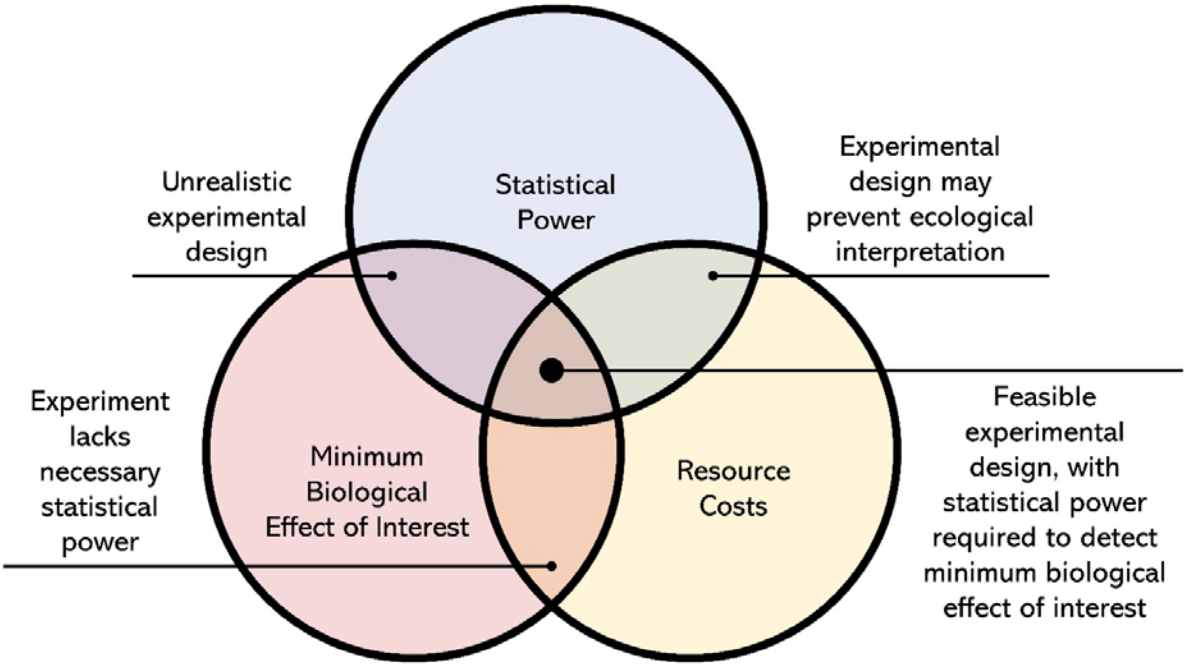

Figure 1. The three considerations important for determining experimental design to investigate how pairs of stressors interact, and the trade-offs that occur when any of them are more limiting than the others.

Stressors: model expectations and interactions

The effects of multiple interacting stressors are commonly determined through the implementation of null models (e.g., Schäfer & Piggott, 2018) where the observed response is compared to an expectation that the stressors are non-interacting (De Laender, 2018). Other methods are available, such as the linear model approach (e.g., Spears et al. 2021), but the issue of sample size is germane to all approaches. Of the range of available null models, the additive null

model (Gurevitch et al., 2000) is the most widely applied (e.g., Crain et al., 2008; Burgess et al. 2021) and has the expectation (null hypothesis) that the overall effect of the multiple interacting stressors is simply equal to the sum of the effects of the stressors acting individually. The statistical test is therefore whether this null model can be rejected in favour of an alternative hypothesis that interactions are: i) greater than anticipated by the additive null model (*Synergistic interactions*); ii) less than the sum of the individual stressor effects (*Antagonistic interactions*); or iii) opposite to that suggested by the additive null model (*Reversal interactions*) (see e.g., Jackson et al., 2016; Orr et al. 2020). The multiplicative null (Lajeunesse, 2011), which is often argued to be preferable for biological responses (e.g., survival) that are bounded (Fournier et al., 2006), is similar to the additive null, except it is additive on a logarithmic scale, meaning it requires all response measures to be greater than zero. This similarity between models, also highlights one of the potential pitfalls in that the user needs to be careful not to log-transform their data before using the additive model and risk the possibility of erroneous interpretation of the results (Griffen et al., 2016). Although we will focus on the additive model as an example to make our case for more experimental replicates, the same issues apply to other null models, and we also show similar results for the multiplicative null model in the supplementary information.

The null model approach requires a factorial experiment design with four treatments that each measure the same biological response metric of interest (e.g., individual survival; population density or biomass; species richness). Each measure X_x, is the mean value of this response metric taken over N_x replicates, where $x \in \{C, A, B, I\}$. The first treatment, *C*, is the control which is simply the system (i.e., individual, population, community) of interest in the absence of any stressor. There are two treatments (*A*, *B*) that account for the response of the system to each of the individual stressors of interest acting in isolation. The final treatment, I, is the estimate of the

response to both stressors acting simultaneously. Associated with each treatment is an estimate of the standard deviation of the response to the treatment, and these are denoted by SD_X, where again $x \in \{C, A, B, I\}$. All three elements, X_x, SD_X, and N_x are required for the additive and multiplicative null models and from this input each null model computes an effect size, with associated confidence intervals from which the interaction type is inferred.

An effect size is a standardized measure of the difference between and expected value and an observed value. We use the definition of effect sizes for factorial experiments under the additive model defined by Gurevitch et al., (2000). The additive null model compares the observed interaction effect $X_O = X_I - X_C$, with $X_E = X_A + X_B - 2X_C$, the expected response (X_E) under the assumption the joint effect is equal to the sum of the individual effects of stressors *A* and *B*. To compute effect sizes, we use Hedges' d which is unbiased by small sample sizes (Hedges & Olkin, 1985). The calculation of the additive effect size, (ES_{Add}), is given as

$$ES_{Add} = \frac{X_E - X_O}{s} \cdot J(m)$$

$$= \frac{X_I - X_A - X_B + X_C}{s} \cdot J(m), \qquad \text{(Equation 1.1)}$$

where s is the pooled standard deviation that takes into account the standard deviations (SD_X) associated with each treatment mean, and $J(m)$ is the small sample bias correction factor (Borenstein et al., 2009). Both of these are defined in the Supplementary Information.

Once computed, we need to know if ES_{Add} is statistically different from 0 in which case the null hypothesis is rejected in favour of an alternative that is dependent on whether ES_{Add} is positive or negative (explored in more detail in the Supplementary Information). This test requires the construction of confidence intervals (at some specified level of statistical significance α), and these

in turn require an estimate of the standard error for our effect size. This first requires an estimate of the variance, estimated as

$$V_{Add} = (J(m))^2 \cdot (\frac{1}{N_I} + \frac{1}{N_A} + \frac{1}{N_B} + \frac{1}{N_C} + \frac{(ES_{Add})^2}{2(N_I + N_A + N_B + N_C)}), \quad \text{(Equation 1.2)}$$

and from this the standard error is computed as

$$SE_{Add} = \sqrt{V_{Add}}, \quad \text{(Equation 1.3)}$$

with the important observation that the standard error for SE_{Add} is *not* divided by the square root of the sample size as is the case for normal estimates of the sampling distribution of a mean. Standard errors should decrease as more samples are taken but increasing sample sizes will already reduce the variance (Equation 1.2), and hence SE_{Add}. Finally, the confidence intervals are computed as

$$CI_{Add} = Z_{\alpha/2} \cdot SE_{Add}, \quad \text{(Equation 1.4)}$$

with $Z_{\alpha/2}$ being the critical Z-score taken at the statistical level of significance α. Typically, α = 0.05, and we divide by two as a two-tailed test is required because the stressors interaction can be less than, or greater than expected under the null model, which means $Z_{\alpha/2} = 1.96$. An important point to note is how the sample sizes N_x appear at multiple stages in the process, with increasing sample sizes leading to smaller confidence intervals for the effect size, and a higher chance that the null hypothesis is rejected (because 0 is not contained within the range covered by the confidence intervals). As the equations contain many terms, it is relatively easy for a small error to creep into the computation of the effect sizes and confidence intervals, although this may be avoided through the use of openly available statistical software (e.g., *multiplestressR*; Burgess & Murrell, 2021).

Resource costs: Typical samples sizes in multiple stressor experiments

Perhaps the most basic question an empirical scientist can ask is "Does my study have sufficient data to answer my question?" (Johnson et al., 2015). In multiple stressor research this amounts to asking whether the sample size is sufficient to detect a departure from the null model of a *given magnitude* should this be the true interaction. We emphasise the qualification of a *given magnitude* as this is where the researcher has to determine *a priori* the smallest deviation from the null expectation that is biologically important. However, this concept has not been discussed, but is critical to knowing how likely we are to be missing important non-null stressor interactions and is a point we focus on in more detail below.

In the absence of any guidance based upon understanding of the null models, researchers have to make sample size decisions that are likely more determined by resource constraints (financial, time, or space costs; Boyd et al., 2018; Rineau et al., 2019), or heuristic arguments (such as the community norm). As a consequence of the lack of statistical guidance, the number of replicates in experiments to investigate stressor interactions rarely reach double figures. For example, two recent meta-analyses (Gomez Isaza et al., 2020; Seifert et al., 2020) included no experiments with more than six replicates per treatment, while a third (Burgess et al., 2021) found <1% of the experiments used more than eight replicates per treatment (Figure 2). Exceptions to this trend tend to focus on individual-level responses with recent examples taken from honeybee health responses to multiple pesticides (Bird et al. 2021) where the mean control treatment sample size was 179.33, and bee responses to pairs of agrochemicals where the mean control treatment sample size for studies where this data is publicly available was 115.62 (Siviter et al. 2021).

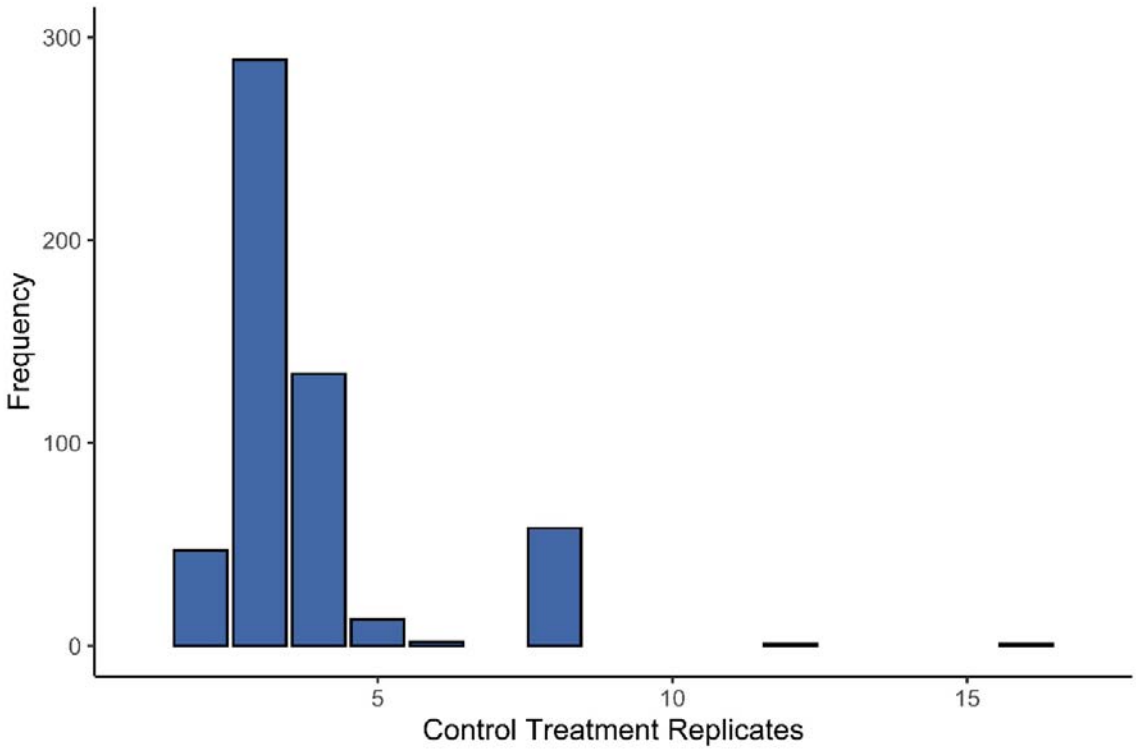

Figure 2. The frequency distribution of control treatment sample sizes from a dataset of 545 stressor interactions in freshwater ecosystems (Burgess et al., 2021).

The importance of sample size for detecting interactions between pairs of co-occurring stressors has only recently been acknowledged. Using a simulated data created from a food web model Burgess et al. (2021) showed how even low levels of observation error, where 99% of all measured responses were within 10% of the true response value, can lead to the inability to detect the true, non-additive interaction in the majority of cases at typical sample sizes of N_x= 4. In other words, even small levels of noise can overwhelm the biological signal when sample sizes are low. With this knowledge, Burgess et al. (2021) were able to conclude that the relatively small proportion (<20%) of non-additive interactions in their freshwater-focussed dataset could easily be explained by the low sample sizes (Figure 2) and that many possibly biologically important non-additive stressor interactions were being missed. However, whilst this warning is useful, it does not answer the question of how many replicates are required.

Critical effect sizes: the smallest detectable interactions

The ability to detect a non-null interaction is dependent on the strength of the interaction, the variation of the biological responses, and the sample sizes (i.e., X_x, SD_X, and N_x), as well as the level of statistical significance α. Both X_x and SD_X, are unknowns and are to be estimated in the experiments, whereas N_x (barring resource costs), and α are both under the control of the researcher. The importance of sample size in detecting non-null interactions can be illustrated with an empirical example (Figure 3). Here, we use the additive null model to determine the effect of stressor pairs on bee health data (Siviter et al., 2021) which comprises a wide range of sample sizes. As expected, increasing sample size results in an increased ability to detect non-null interactions, and we can see how greater sample sizes allow weaker non-null interactions to be identified and classified (Figure 3).

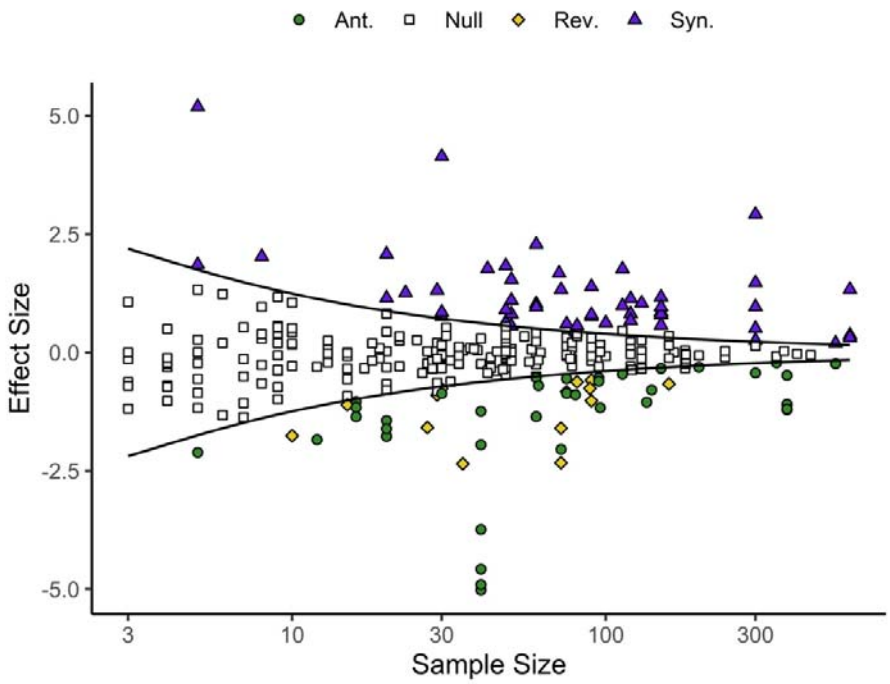

Figure 3. The effect of sample size on the ability to detect interactions with different effect sizes for the bee health responses to multiple stressors in Siviter et al. (2021). Open squares denote data points that are statistically indistinguishable from the null model of an additive interaction (i.e., the null model that co-occurring stressors are simply the sum of their individual effects). Data points that lead to the rejection of the null model can be assigned as synergistic (blue triangles), antagonistic (green circles), or reversals (orange diamonds). The black lines denote the critical effect size that separates the region of detectable departure from the null model at the 5% level of significance. Median sample size per treatment is plotted on the x axis. A small number of null interactions appear outside of the null region where the experiment had uneven sample sizes between treatments, but for clarity of presentation the critical effect size is computed under the assumption of equal sample sizes within each study. Figure 3 was generated using the *multiplestressR* R package (Burgess & Murrell, 2021), with code to replicate this figure provided in the Supplementary Material.

For each sample size, there is a minimum effect size which any experiment will be able to distinguish as being statistically different to the null model (illustrated by the black lines in Figure 3). Effect sizes below this threshold denote interactions that cannot be distinguished from the null model expectation of additivity at the chosen level of statistical significance. This threshold, referred to as the *Critical Effect Size* (see Mudge et al., 2012; Lakens, 2021) can be exactly calculated for the additive null model (the equation for which is detailed in the Supplementary Information but can be computed using the R library *multiplestressR*; Burgess & Murrell, 2021). Analysis of the bee health data (Siviter et al., 2021) shows how the critical effect size (ES_{Add}) predicts non-additive interactions and verifies the expectation that only very large effect sizes can reject the null expectation of additivity when sample sizes are below 20 per treatment (Figure 3). A similar, but messier relationship holds for the multiplicative null model, albeit with a different scale on the y-axis, and we detail this in the Supplementary Information. Unfortunately, the bee stressor response dataset from Siviter et al. (2021) is very much an outlier when compared to most other datasets (e.g., Figure 2). At the very low samples sizes that typify multiple stressor research, especially for population- and community-level responses (e.g., Figure 2), effect sizes have to be very large (e.g., for $N_x = 4, ES_{Add} \sim 2$) in order for non-additive interactions to be detected.

Smallest interaction of interest: *What is a biologically meaningful interaction?*

The critical effect size states the smallest effect size that can be detected in a study with given sample sizes for the treatments and chosen level of statistical significance. However, the critical effect size relates to *statistical* but *not biological* significance. On the one hand, small sample sizes can lead to the detection of only large effect sizes (Figure 3), but at the other end of the scale infinitely large sample sizes can detect infinitely small departures from additivity (i.e., the lines in

Figure 3 asymptote slowly to 0). So, whilst small sample sizes likely miss key stressor interactions, unnecessarily large sample sizes can waste resources (Figure 1). To avoid either of these outcomes, the researcher needs to determine the smallest interaction that would lead to a biologically meaningful deviation from the null model before the experiment is run (to avoid any bias from knowing the result). We define this interaction as the minimum biological effect size, and we argue this depends upon both the study system and response of interest. For example, a researcher may want to determine whether two stressors combine to affect a response (e.g., juvenile survival rates) in a non-additive manner for an endemic or threatened species. In this scenario, it is important to be able to detect a small deviation from additivity (i.e., a small effect size), as failing to detect even a weak interaction may lead to the wrong mitigation strategy being selected and potentially exacerbate the effects of these stressors to the detriment of the study system (Brown et al., 2013; Côté et al., 2016). Commonplace sample sizes (e.g., ~4 replicates per treatment) are not adequate for this question (Figure 3), and the researcher will likely need to implement sample sizes which are multiple (i.e., two or more) orders of magnitude larger than those commonly used. There may be other situations where a smaller effect is not so important, implying smaller samples are adequate, such as monitoring abundance declines in a system with high functional redundancy, but even here care needs to be taken since concerns have been raised regarding the overestimation of stressor effects from experiments with small sample sizes (Yang et al., 2021).

How should the minimum effect size of interest be determined? Although it might seem tempting to use the heuristic guidelines proposed by Cohen (1988) for small, medium, large effect sizes, we do not believe they are appropriate for multiple stressor research due to the heterogeneity in systems, responses, and stressors. For example, would we decide upon the same minimum effect

size for survival responses at different stages in a species' life cycle? In any case these guidelines only relate to Cohen's *d* or Hedge's *d* and do not apply to null models such as the multiplicative null model that operate on a different scale. Other ways that the minimum effect size of biological interest could be determined include guidance from ecological theory, and results of previous meta-analyses (Lakens 2021). However, in order for a theoretical model to be a useful guide, it needs to be an adequate approximation to the stressors, biological system, and response under scrutiny. This is a tall ask, since it is likely that empirical evidence is required to calibrate the model in the first place, in which case there is already some evidence that could be used (carefully) to consider the number of replicates required. The results of previous meta-analyses could act as a guide, although again care needs to be taken since it is possible that publication biases towards biologically novel but not necessarily statistically robust effect sizes (Filazzola and Cahill Jr. 2021) could affect summary effect sizes. Moreover, meta-analyses in ecology and evolution often report high levels of heterogeneity (Senior et al., 2016) compared to clinical trials as ecological and evolutionary studies often focus on multiple taxa, in real world environments, that are subject to many different forms of environmental and biological variation (Burgess et al. 2021; Côté et al., 2016). It is therefore hard to know if the summary effect sizes reported in these studies are relevant for other, more focussed, studies that might being asking subtly different questions involving, for example, different stressors or responses.

Consequences and recommendations

Overall, we do not believe there is a simple answer to the smallest effect size of biological interest. Instead, researchers need to consider it explicitly in the light of their question, system, and responses. This could be done by using any prior knowledge to arrive at likely quantities for X_x, and SD_X, and then determining what the smallest difference between observed and expected

interaction (Equation 1.1) should be. This could then be used to determine the desired sample size using the critical effect size equation (see *multiplestressR;* Burgess & Murrell, 2021). The sweet-spot of sample size is dependent on the trifecta of resource costs, statistical power, and minimum effect of biological interest, and failure to take any of these into consideration may limit the effectiveness of any experiment (Figure 1). However, it seems likely that in many cases $N_x = 4$ does in fact lead to biologically important on-null stressor-pair interactions being left undetected and given the relationship between critical effect size and sample size, 20 replicates (or more) might be desirable.

The recent meta-analyses of how pairs of pesticides interact to affect bee health (Siviter et al. 2021, Bird et al. 2021) are examples of experiments with very large sample sizes, and the fact that they both focus on studies at the individual-level highlight how this might be the resource efficient way to increase replicate numbers. This echoes earlier calls to focus on individual-level responses to stressors as it is the fate and/or behaviour of the individual that is directly affected (e.g., Maltby 1999). However, responses at other (higher) levels of biological complexity such as population, community and ecosystem are also likely to be of interest because it is the response of these levels that may matter the most from a stressor management standpoint (Simmons et al. 2021). Moreover, because each species is embedded within a food web, interactions between species can lead to compensatory (antagonism) or synergistic effects that are not observed for individual species in isolation (Christensen et al., 2006; Burgess et al. 2021; Simmons et al. 2021). Unfortunately, it is much harder to increase the sample sizes of many mesocosm experiments for these higher levels of organisation simply due to the financial cost, space, and time required to manage large sample sizes for all four treatments (Boyd et al., 2018). One alternative to boost within-study replication is to use coordinated networks of researchers who ask the same

experimental question(s) across multiple sites, using the same protocol (Filazzola and Cahill Jr. 2021; Yang et al., 2021). An example of this is the Nutrient Network (NutNet) organisation (https://nutnet.org/) that amongst its key questions asks: To what extent are plant production and diversity co-limited by multiple nutrients in herbaceous-dominated communities? Another instance of this linked approach is the Managing Aquatic ecosystems and water resources under multiple stress (MARS) project (Hering et al., 2015) that has investigated the responses of a large number of European water bodies to multiple stressors (e.g., Birk et al. 2020). As always, there is no silver bullet, and coordinated networks may suffer from increases in data heterogeneity due to the multiple site nature of the network and the natural environmental and biological variation this includes, but also because small, but important differences in protocol may occur simply due to the number of research teams implementing the framework (Filazzola and Cahill Jr. 2021).

So far, our discussions of null models and sample sizes have generally been restricted to investigations of pairs of stressors, yet we know that many ecosystems are being challenged with more than two stressors (Halpern et al., 2015). For example, Nõges et al. (2016) identified European waters with up to seven co-acting stressors, although two co-acting stressors were the most common, being identified in 42% of cases. Similarly, there have been calls for investigating the responses to stressors at multiple levels of intensity (Polazzo et al. 2021; Schäfer & Piggott, 2018), since responses at low and high stressor intensities may differ greatly (Beaumelle et al., 2020; Dixon et al., 2020) and result in different interactions being detected (Ma et al., 2020). In both cases, sample sizes will need to be even larger than for two stressors each at a single intensity, and as we have already found, many experiments are probably greatly underpowered even in this simpler scenario (Yang et al. 2021). In order to maximise the outcome for the input of resources we suggest that individual studies should first try to boost sample sizes for simpler

experiments before adding in further complexity, and encourage investigations of greater than two stressors and/or multiple intensities to use coordinated networks where the sample sizes can be distributed across multiple research teams, or focus on individual-level responses where sample sizes may more easily run into the hundreds (e.g. Bird et al. 2021; Siviter et al. 2021).

Ultimately, resource constraints, especially for studies on responses at higher levels of biological organisation, may mean it is not possible to design an experiment with adequate sample sizes to capture biologically interesting/important stressor-pair interactions. However, such experiments can still be useful as they provide data for meta-analyses that collate individual experiments together to greatly increase the power to correctly reject the null model (e.g., Crain et al., 2008; Jackson et al., 2016; Przeslawski et al., 2015). The key point is that to aid general understanding, and avoid publication bias, it is crucial that all experiments are published with the data made openly available (i.e., the three components of sample size, mean and standard deviation/error or variance for each treatment) and not just those experiments that detect 'interesting' non-null stressor-pair interactions (Filazzola and Cahill Jr.2021). Indeed, it is likely that publication bias is leading to the effects of anthropogenic stressors being overestimated (Yang et al., 2021), while multiple stressor ecology suffers from the erroneous over-reporting of synergistic interactions (Côté et al., 2016). Unfortunately, there are still many papers that do not report or make their data (i.e., treatment means etc.) readily available. For example, Burgess et al. (2021) identified 122 papers that appeared suitable for their meta-analysis of freshwater stressor interactions, but 66 had to be discarded due to missing data or having figures that were too unclear for data extraction. Not reporting these data represents a waste of resources, as it prevents future analyses (which are often unanticipated during the original study) from being conducted (Hanson & Walker, 2020).

Conclusions

Our aim here was to open the discussion regarding sample sizes in multiple stressor research and show that before we ask the question "how much data do I need?", we first need to answer the question "what is a biologically important interaction?". Increasing sample sizes will always lead to an improvement in our statistical ability to detect unexpected stressor-pair interactions, but at extreme sample sizes we will likely be detecting only very small departures from the null model and these may not necessarily be relevant for management decisions. Setting the lower bound for an interesting stressor-pair interaction is critical to knowing what sample sizes are required. Currently, it is our view that many experiments are likely underpowered and missing biologically important interactions, but studies that mostly focus on individual-level responses to stressors may be more adequately sampled. Strategies such as research networks may help increase sample sizes for higher levels of biological organisation such as communities, but there is still potential value in conducting smaller-scale studies, as long as they are published and data made freely available, since they can contribute to meta-analyses and aid the design of subsequent experiments. Finally, although we have focussed on the commonly used additive and multiplicative null models, there are a number of other null models that have been proposed (e.g., Schäfer & Piggott, 2018; Dey & Koops, 2021), and to date there is no guidance on sample sizes required to detect non-null interactions of any given magnitude. This needs to be remedied. Until we can quantify the abilities of the statistical models to detect different strengths of interactions, we will be kept in the dark about how many unexpected interactions we are missing, and the amount of data required to uncover them.

Acknowledgments

This research was supported by the United Kingdom Natural Environment Research Council (NERC) via the studentship grant NE/M010481/1, and the research grant NE/V001396/1. The authors declare no conflicts of interest.

Author contributions

BB and DM performed the analyses and drafted the manuscript. BB derived the critical effect size for the additive null model. All the authors contributed significantly to the intellectual core of the manuscript; to the interpretation of the results; and to revisions of the manuscript.

Data availability

All data analysed within this paper is openly available. Code to generate Figure 3 is provided in the Supplementary Material.

References

Ban, S. S., Graham, N. A., & Connolly, S. R. 2014. Evidence for multiple stressor interactions and effects on coral reefs. *Global Change Biology*, 20(3), 681– 697.

Beauchesne, D., Cazelles, K., Archambault, P., Dee, L., & Gravel, D. 2021. On the sensitivity of food webs to multiple stressors. *Ecology Letters*, 24, 2219– 2237. https://doi.org/10.1111/ele.13841

Beaumelle, L., De Laender, F., & Eisenhauer, N. 2020. Biodiversity mediates the effects of stressors but not nutrients on litter decomposition. *Elife*, *9*, e55659.

Bird, G., Wilson, A. E., Williams, G.R., Hardy, N.B. 2021. Parasites and pesticides act antagonistically on honey bee health. *Journal of Applied Ecology*, 58: 997– 1005. https://doi.org/10.1111/1365-2664.13811

Borenstein, M., Cooper, H., Hedges, L. and Valentine, J., 2009. Effect sizes for continuous data. *The Handbook of Research Synthesis and Meta-Analysis*, 2, pp.221-235.

Boyd, P. W., Collins, S., Dupont, S., Fabricius, K., Gattuso, J. P., Havenhand, J., ... & Pörtner, H. O. 2018. Experimental strategies to assess the biological ramifications of multiple drivers of global ocean change—a review. Global Change Biology, 24(6), 2239-2261.

Brown, C.J., Saunders, M.I., Possingham, H.P. and Richardson, A.J., 2013. Managing for interactions between local and global stressors of ecosystems. *PLoS One*, 8(6), p.e65765.

Burgess B. J. and Murrell D. J., 2021. multiplestressR: Additive and Multiplicative Null Models for Multiple Stressor Data. R package version 0.1.1. https://CRAN.R-project.org/package=multiplestressR

Burgess, B.J., Purves, D., Mace, G. and Murrell, D.J., 2021. Classifying ecosystem stressor interactions: Theory highlights the data limitations of the additive null model and the difficulty in revealing ecological surprises. *Global Change Biology*, 27: 3052-3065. https://doi.org/10.1111/gcb.15630

Christensen, M.R., Graham, M.D., Vinebrooke, R.D., Findlay, D.L., Paterson, M.J. and Turner, M.A., 2006. Multiple anthropogenic stressors cause ecological surprises in boreal lakes. *Global Change Biology*, 12(12), pp.2316-2322.

Cohen, J. (1988). Statistical Power Analysis for the Behavioral Sciences (2nd ed.). Routledge. https://doi.org/10.4324/9780203771587.

Côté, I.M., Darling, E.S. and Brown, C.J., 2016. Interactions among ecosystem stressors and their importance in conservation. *Proceedings of the Royal Society B: Biological Sciences*, *283*(1824), p.20152592.

Crain, C.M., Kroeker, K. and Halpern, B.S., 2008. Interactive and cumulative effects of multiple human stressors in marine systems. *Ecology Letters*, 11(12), pp.1304-1315.

De Laender, F., 2018. Community-and ecosystem-level effects of multiple environmental change drivers: Beyond null model testing. *Global Change Biology*, 24(11), pp.5021-5030.

Dey, C.J. and Koops, M.A., 2021. The consequences of null model selection for predicting mortality from multiple stressors. *Proceedings of the Royal Society B*, 288(1948), p.20203126.

Dixon, G., Abbott, E., & Matz, M. 2020. Meta-analysis of the coral environmental stress response: Acropora corals show opposing responses depending on stress intensity. *Molecular Ecology*, *29*(15), 2855-2870.

Filazzola, A., & Cahill, J. F. 2021. Replication in field ecology: Identifying challenges and proposing solutions. *Methods in Ecology and Evolution*, 12, 1780– 1792. https://doi.org/10.1111/2041-210X.13657

Fournier, V., Rosenheim, J.A., Brodeur, J., Diez, J.M. and Johnson, M.W., 2006. Multiple plant exploiters on a shared host: testing for nonadditive effects on plant performance. *Ecological Applications*, 16(6), pp.2382-2398.

Gomez Isaza, D.F., Cramp, R.L. and Franklin, C.E., 2020. Living in polluted waters: a meta-analysis of the effects of nitrate and interactions with other environmental stressors on freshwater taxa. *Environmental Pollution*, p.114091.

Griffen, B. D., Belgrad, B. A., Cannizzo, Z. J., Knotts, E. R., & Hancock, E. R. 2016. Rethinking our approach to multiple stressor studies in marine environments. *Marine Ecology Progress Series*, 543, 273-281.

Gurevitch, J., Morrison, J.A. and Hedges, L.V., 2000. The interaction between competition and predation: a meta-analysis of field experiments. *The American Naturalist*, 155(4), pp.435-453.

Halpern, B. S., Frazier, M., Potapenko, J., Casey, K. S., Koenig, K., Longo, C., ... & Walbridge, S. 2015. Spatial and temporal changes in cumulative human impacts on the world's ocean. Nature Communications, 6(1), 1-7.

Hanson, P. J., & Walker, A. P. 2020. Advancing global change biology through experimental manipulations: Where have we been and where might we go?. *Global change biology*, *26*(1), 287-299.

Hedges, L. V., and I. Olkin., 1985. Statistical methods for meta-analysis. *Academic Press*, New York.

Hering, D., Carvalho, L., Argillier, C., Beklioglu, M., Borja, A., Cardoso, A. C., ... & Birk, S. 2015. Managing aquatic ecosystems and water resources under multiple stress—An introduction to the MARS project. *Science of the total environment*, *503*, 10-21.

Hodgson, E.E. and Halpern, B.S., 2018. Investigating cumulative effects across ecological scales. *Conservation Biology*, 33(1), pp.22-32.

Jackson, M.C., Loewen, C.J., Vinebrooke, R.D. and Chimimba, C.T., 2016. Net effects of multiple stressors in freshwater ecosystems: a meta-analysis. *Global Change Biology*, 22(1), pp.180-189.

Jackson, M.C., Pawar, S. and Woodward, G., 2021. The Temporal Dynamics of Multiple Stressor Effects: From Individuals to Ecosystems. *Trends in Ecology & Evolution,* 36(5): 402-410.

Johnson, P.C., Barry, S.J., Ferguson, H.M. and Müller, P., 2015. Power analysis for generalized linear mixed models in ecology and evolution. *Methods in Ecology and Evolution*, 6(2), pp.133-142.

Lajeunesse, M.J., 2011. On the meta-analysis of response ratios for studies with correlated and multi-group designs. *Ecology*, 92(11), pp.2049-2055.

Lakens, D., 2021. Sample Size Justification. *PsyArXiv*. doi:10.31234/osf.io/9d3yf.

Lange, K., Bruder, A., Matthaei, C.D., Brodersen, J. and Paterson, R.A., 2018. Multiple-stressor effects on freshwater fish: Importance of taxonomy and life stage. *Fish and Fisheries*, 19(6), pp.974-983.

Lemm, J. U., Venohr, M., Globevnik, L., Stefanidis, K., Panagopoulos, Y., van Gils, J., ... & Birk, S. 2021. Multiple stressors determine river ecological status at the European scale: Towards an integrated understanding of river status deterioration. *Global Change Biology*, *27*(9), 1962-1975.

Ma, Z., Chen, H. Y., Li, Y., & Chang, S. X. 2020. Interactive effects of global change factors on terrestrial net primary productivity are treatment length and intensity dependent. *Journal of Ecology*, *108*(5), 2083-2094.

Maltby, L. 1999. Studying stress: The importance of organism-level responses. *Ecological Applications*, 9(2), 431-440.

Mudge, J.F., Baker, L.F., Edge, C.B. and Houlahan, J.E., 2012. Setting an optimal α that minimizes errors in null hypothesis significance tests. *PLoS One*, 7(2), p.e32734.

Nõges P, Argillier C, Borja Á, Garmendia JM, Hanganu J, Kodeš V, Pletterbauer F, Sagouis A, Birk S. 2016. Quantified biotic and abiotic responses to multiple stress in freshwater, marine and ground waters. Science of the Total Environment. 540: 43-52. doi: 10.1016/j.scitotenv.2015.06.045. Epub 2015

Orr, J.A., Vinebrooke, R.D., Jackson, M.C., Kroeker, K.J., Kordas, R.L., Mantyka-Pringle, C., Van den Brink, P.J., De Laender, F., Stoks, R., Holmstrup, M. and Matthaei, C.D., 2020. Towards a unified study of multiple stressors: divisions and common goals across research disciplines. *Proceedings of the Royal Society B*, 287(1926), p.20200421.

Polazzo, F., Roth, S. K., Hermann, M., Mangold-Döring, A., Rico, A., Sobek, A., ... & Jackson, M. C. 2021. Combined effects of heatwaves and micropollutants on freshwater ecosystems: Towards an integrated assessment of extreme events in multiple stressors research. *Global change biology* 00, 1– 20. https://doi.org/10.1111/gcb.15971.

Przeslawski, R., Byrne, M., & Mellin, C. 2015. A review and meta-analysis of the effects of multiple abiotic stressors on marine embryos and larvae. *Global Change Biology*, 21(6), 2122-2140.

Rineau, F., Malina, R., Beenaerts, N., Arnauts, N., Bardgett, R. D., Berg, M. P., ... & Vangronsveld, J. 2019. Towards more predictive and interdisciplinary climate change ecosystem experiments. *Nature climate change*, *9*(11), 809-816.

Schäfer, R.B. and Piggott, J.J., 2018. Advancing understanding and prediction in multiple stressor research through a mechanistic basis for null models. *Global Change Biology*, 24(5), pp.1817-1826.

Seifert, M., Rost, B., Trimborn, S., & Hauck, J. 2020. Meta-analysis of multiple driver effects on marine phytoplankton highlights modulating role of pCO2. *Global Change Biology*, 26(12), 6787-6804.

Senior, A. M., Grueber, C. E., Kamiya, T., Lagisz, M., O'Dwyer, K., Santos, E. S., & Nakagawa, S. 2016. Heterogeneity in ecological and evolutionary meta-analyses: its magnitude and implications. *Ecology*, 97(12), 3293–3299.

Simmons, B.I., Blyth, P.S.A., Blanchard, J.L. *et al.* 2021. Refocusing multiple stressor research around the targets and scales of ecological impacts. *Nature Ecology and Evolution,* 5: 1478–1489 https://doi.org/10.1038/s41559-021-01547-4.

Siviter, H., Bailes, E. J., Martin, C. D., Oliver, T. R., Koricheva, J., Leadbeater, E., & Brown, M. J. 2021. Agrochemicals interact synergistically to increase bee mortality. Nature, 596(7872), 389-392.

Sokolova, I., 2021. Bioenergetics in environmental adaptation and stress tolerance of aquatic ectotherms: linking physiology and ecology in a multi-stressor landscape. *Journal of Experimental Biology*, 224(Suppl_1). doi:10.1242/jeb.236802.

Spears, B. M., Chapman, D. S., Carvalho, L., Feld, C. K., Gessner, M. O., Piggott, J. J., ... & Birk, S. 2021. Making Waves. Bridging theory and practice towards multiple stressor management in freshwater ecosystems. *Water Research*, 116981.

Yang, Y., Hillebrand, H., Lagisz, M., Cleasby, I., & Nakagawa, S. 2021. Low statistical power and overestimated anthropogenic impacts, exacerbated by publication bias, dominate field studies in global change biology. *Global change biology*, 28, 969– 989. https://doi.org/10.1111/gcb.15972.

Supplementary Information

Additive null model

Here we give the full details of the additive null model outlined by Gurevitch et al., (2000). The calculation of the additive effect size, (ES_{Add}), is given by Equation S1.1, and this requires a pooled sampling standard deviation, denoted by *s* (Equation S1.2); while *J(m)* (Equation S1.3) is the small sample bias correction factor, often used where the there is a small number of replicates per interaction (Borenstein et al., 2009). For each additive effect size a corresponding variance, (V_{Add}), standard error, (SE_{Add}), and confidence intervals, (CI_{Add}), are calculated by Equations S1.4 - S1.6. Here, we set the significance level to 0.05 ($Z_{\alpha/2}$ = 1.96), hence 95% confidence intervals are calculated.

$$S1.1) \qquad ES_{Add} = \frac{X_I - X}{}$$

$$\cdot J(m)$$

$$S1.2) \quad s$$

$$= \sqrt{\frac{(N_I - 1) \cdot (SD_I)^2 + (N_A - 1) \cdot (SD_A)^2 + (N_B - 1) \cdot (SD_B)^2 + (N_C -}{N_I + N_A + N_B + N_C - 4}}$$

$$S1.3) \qquad J(m) = 1 - \frac{3}{4 \cdot (N_I + N_A + N_B + N_C - 4) - 1}$$

$$S1.4) \qquad V_{Add} = (J(m))^2 \cdot (\frac{1}{N_I} + \frac{1}{N_A} + \frac{1}{N_B} + \frac{1}{N_C} + \frac{(ES_{Add})^2}{2(N_I + N_A + N_B + N_C)})$$

$$S1.5) \qquad SE_{Add} = \sqrt{V_{Add}}$$

$$S1.6) \qquad CI_{Add} = Z_{\alpha/2} \cdot SE_{Add}$$

This is the version implemented in the data analysis in Figure 3 of the main text.

Multiplicative null model

We implement the form of the multiplicative null model detailed by Lajeunesse (2011). The multiplicative effect size, (ES_{Mul}), is calculated by Equation S2.1; while the corresponding variance, (V_{Mul}), standard error, (SE_{Mul}), and confidence intervals, (CI_{Mul}), are determined using Equations S2.2 – S2.4. As above, we set the significance level to 0.05 ($Z_{\alpha/2}$ = 1.96), hence 95% confidence intervals are calculated.

$$S_{Mul} = \ln(X_I) - \ln(X_A) - \ln(X_B) + \ln(X_C) \quad S2.1)$$

$$V_{Mul} = \frac{(SD_I)^2}{(X_I)^2 \cdot N_I} + \frac{(SD_A)^2}{(X_A)^2 \cdot N_A} + \frac{(SD_B)^2}{(X_B)^2 \cdot N_B} + \frac{(SD_C)^2}{(X_C)^2 \cdot N_C} \quad S2.2)$$

$$SE_{Mul} = \sqrt{V_{Mul}} \quad S2.3)$$

$$CI_{Mul} = Z_{\alpha/2} \cdot SE_{Mul} \quad S2.4)$$

We draw the reader's attention to the observation that, in contrast to the additive null model (Equation S1) the variance for the multiplicative model (Equation S2.2) is independent of the effects size.

Interaction classifications

As described in the main text interaction can be classified as a synergistic, antagonistic, reversal, or null interaction using either the additive (Gurevitch et al., 2000) or multiplicative (Lajeunesse, 2011) null models. As detailed in the *Methods*, the classification of an interaction is dependent upon the effect size (ES), associated confidence intervals ($CI_{95\%}$), the observed ($Obs.$) and predicted ($Pred.$) interactive effect of stressors.

For the additive null model, the $Pred.$ can be calculated as $X_A + X_B - 2X_C$; while for the multiplicative null model, $Pred.$ can be calculated as $\ln\left(\frac{X_A X_B}{X_C^2}\right)$. Similarly, for the additive null model, $Obs.$ can be calculated as $X_I - X_C$; while for the multiplicative null model, $Obs.$ Can be calculated as $\ln\left(\frac{X_I}{X_C}\right)$. Using this scheme, interactions were classified as follows:

When $Pred. \geq 0$

1) If $ES \pm CI_{95\%}$ overlaps zero, then the interaction was assigned a null classification.
2) If $ES > 0$, the interaction was assigned a synergistic classification.
3) If $ES < 0$, and if $Obs.$ and $Pred.$ both have the same polarity, the interaction was assigned an antagonistic classification.
4) If $ES < 0$, and if $Obs.$ and $Pred.$ have contrasting polarities, the interaction was assigned a reversal classification.

When $Pred. < 0$

1) If $ES \pm CI_{95\%}$ overlaps zero, then the interaction was assigned a null classification.
2) If $ES < 0$, the interaction was assigned a synergistic classification.
3) If $ES > 0$, and if $Obs.$ and $Pred.$ both have the same polarity, the interaction was assigned an antagonistic classification.
4) If $ES > 0$, and if $Obs.$ and $Pred.$ have contrasting polarities, the interaction was assigned a reversal classification.

Critical Effect Size

By combining and rearranging Equations S1.1 – S1.6, it is possible to express the additive effect size as a function of sample size:

$$|ES_{Add}| > \sqrt{\left(\frac{1}{N_C} + \frac{1}{N_A} + \frac{1}{N_B} + \frac{1}{N_I}\right) \cdot \left(\frac{2 \cdot (Z_{\alpha/2})^2 \cdot (N_C + N_A + N_B + N_I) \cdot \left(J(m)\right)^2}{2 \cdot (N_C + N_A + N_B + N_I) - (Z_{\alpha/2})^2 \cdot \left(J(m)\right)^2}\right)}$$

(Inequality S1)

This inequality shows that for a given sample size, there is a minimum additive effect size value (also been referred to as the critical effect size by Lakens 2021) which must be exceeded before the interaction can be correctly classified as being non-null.

Unfortunately, due to a more complex structure in the calculation of the variance the multiplicative null model does not yield such an inequality for calculating the critical effect size. However, we repeat the data analysis shown in Figure 3, but using the multiplicative null model instead of the additive null model (Figure S1). Although the relationship is much less straightforward there appears to be a higher probability of rejecting the null model expectation in favour of one of the three alternative interactions (antagonistic, reversal, synergistic).

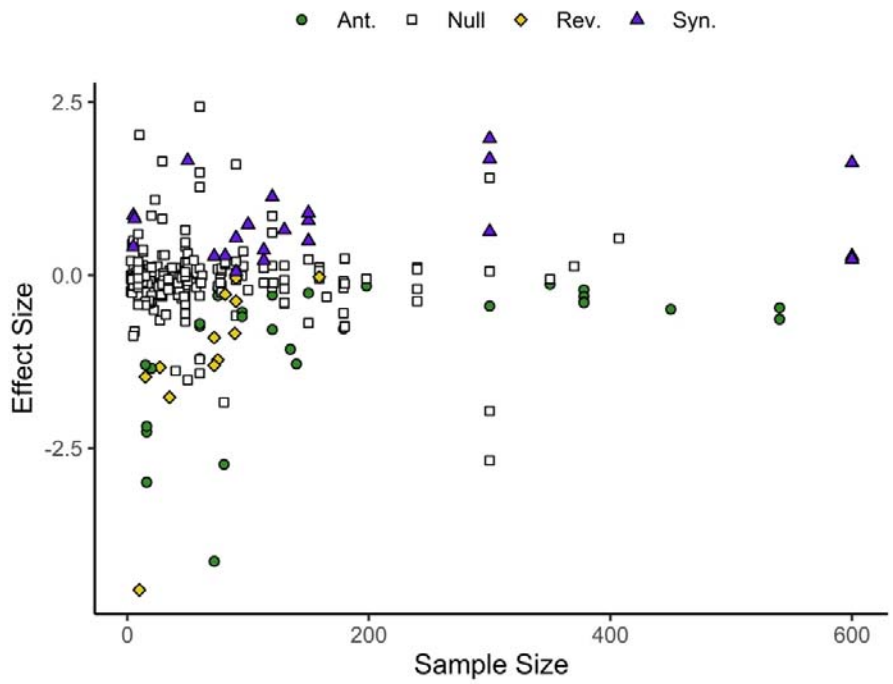

Figure S1. The effect of sample size on the ability to detect interactions with different effect sizes for the bee health responses to multiple stressors in Siviter et al. (2021). Open squares denote data points that are statistically indistinguishable from the null model of a multiplicative interaction (Equation S2). Data points that lead to the rejection of the null model can be assigned as synergistic (blue triangles), antagonistic (green circles), or reversals (orange diamond). Median sample size per treatment is plotted on the x axis. This figure was generated using the *multiplestressR* R package (Burgess & Murrell, 2021).

References

Borenstein, M., Cooper, H., Hedges, L. and Valentine, J., 2009. Effect sizes for continuous data. *The Handbook of Research Synthesis and Meta-Analysis*, 2, pp.221-235.

Burgess B. J. and Murrell D. J., 2021. multiplestressR: Additive and Multiplicative Null Models for Multiple Stressor Data. R package version 0.1.1. https://CRAN.R-project.org/package=multiplestressR.

Gurevitch, J., Morrison, J.A. and Hedges, L.V., 2000. The interaction between competition and predation: a meta-analysis of field experiments. *The American Naturalist*, 155(4), pp.435-453.

Lajeunesse, M.J., 2011. On the meta-analysis of response ratios for studies with correlated and multi-group designs. *Ecology*, 92(11), pp.2049-2055.

Lakens, D., 2021. Sample Size Justification. *PsyArXiv*. doi:10.31234/osf.io/9d3yf.

Siviter, H., Bailes, E. J., Martin, C. D., Oliver, T. R., Koricheva, J., Leadbeater, E., & Brown, M. J. 2021. Agrochemicals interact synergistically to increase bee mortality. *Nature*, 596(7872), 389-392.