Enhancing understanding of experimental designs: treatment levels and choice of analytics to improve statistical performance for ecological experiments

Justin Cuffaro

A Thesis in the Department of

Biology

Presented in Partial Fulfillment of the Requirements

For the Degree of Master of Science (Biology)

At Concordia University

Montréal, Québec, Canada

September 2023

This is to certify that the thesis prepared

By: Justin Cuffaro

Entitled: Enhancing understanding of experimental designs: treatment levels and choice of analytics to improve statistical performance for ecological experiments

and submitted in partial fulfillment of the requirements for the degree of

Master of Science (Biology)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

_____ Chair

Dr. Emma Despland

_____ Examiner

Dr. Grant Brown

_____ External Examiner

Dr. Eric Pedersen

_____ Co-Supervisor

Dr. James Grant

_____ Supervisor

Dr. Pedro Peres-Neto

Approved by

_____

Dr. Robert Weladji, Graduate Program Director

Date: September 6th, 2023

_____

Dr. Pascale Sicotte, Dean of Faculty

Enhancing understanding of experimental designs: treatment levels and choice of analytics to improve statistical performance for ecological experiments

Justin Cuffaro

Abstract

Experimental design is a fundamental component of research in ecology and other disciplines. It is critical to understand the consequences of statistical inference, including power and effect size, when making decisions about designing experiments. However, issues such as file drawer effect, funding logistics, and reproducibility are a major concern that often are not considered when starting one's scientific journey; a poor understanding of these problems may lead to overly conservative estimates or claims that can not be replicated. Here we argue that researchers can dramatically improve inferences from experiments by focusing on two key issues. First, properly manipulating treatment dispersion, which refers to the variation among levels of a quantitative factor, can improve inference without the need for increasing replicates and sample sizes. Secondly, choosing analytics judiciously, such as selecting between ANOVA and replicated regression for experimental data, can improve inference by contrasting inferential outcomes on the same data. We use language, simple fictional examples, and simulations to show that effect size and power increase with treatment dispersion. We also conducted a small meta-analysis on real data to assess whether the literature confirms in published data that treatment dispersion affects inference. We found that there is no association between treatment dispersion and effect size in published literature, suggesting that some form of bias may be prevalent in published literature. Overall, we found that by focusing on treatment dispersion and analytics, researchers can improve their ability to make sound inferences from their data without the need for increased sample sizes.

Acknowledgements

The wide sweeping and global effect of the Covid19 pandemic has had major consequences on us all and we've all had to adapt in some way. It made us afraid to leave the house and to interact with each other. It promoted isolation and forced us to take a deep look onto who we are as a society. It has changed me profoundly, as I'm sure it has all of you. However, despite the struggle, despite the battle that is each day, I cannot help but look at all the positives that have been and that have come into my life. I started a new job (I bit crazy to do while writing a thesis I know) and I've learned a lot from my coworkers and friends about tackling new challenges head on. I moved out for the first time, alone. There was a certain challenge to it, but I realized there was a certain freedom and satisfaction of being the sole person responsible for your well being. I dove headfirst into my project, and I learned a lot about statistics and the models we as biologists make sense of the data in front of us. I met many good friends in Pedro's lab, and since I stayed a lot longer than I should have, I had the pleasure to meet and witness multiple iterations that I thank dearly for their support. I especially thank Gabriel for putting up with me sharing his desk, as well helping me solve multiple problems while working on this thesis. I want to thank my partner of 2 years Karine, for being there with me throughout, and bringing light into my world at the end of the day. Coming home is always better when you're there. I want to thank my parents and my brother, for supporting me throughout my entire life. Especially my mom (who mentioned me in her thesis as a 1-year-old) who always makes the best lasagna. It goes without saying that I want to thank Pedro and Jim for their support and guidance on this thesis, I've learned so much since I've started, and I hope I've made you proud. I also want to thank NSERC-DG to Pedro and Concordia-FAS for the support that they've given us throughout. A special shoutout to the Fishmen, you know who you are. The last acknowledgement goes to you, the reader, I hope you learn as much as I have, and find the information useful for your meta-analysis (or other endeavors).

Table of Contents

List of Figures

List of Tables

Introduction

Scientific research is often communicated in the form of journals featuring published articles. While there exists a robust peer review system in place to make sure published articles adhere to the scientific method and are of high quality, the system itself is often slightly controversial and subject to debate (Ware 2008). Likewise, the system fails to address problems such as file drawer effect and reproducibility of the experiment. File drawer effect is simply the fact that papers with low effect size or non-significant results tend to be unpublished, locked in the file drawer so to say (Scargle 1999). File drawer effect is especially prevalent as reviewers are more inclined to publish high impact, high effect size articles, because impact factor is a huge part of a journal's reputation (Dong et al. 2005). Reproducibility is another problem that is highly relevant today. Reproducibility is the ability of an experiment to be replicated with the exact same methodology by (though not necessarily) a third party. Reproducibility is a huge consideration when looking at the value of scientific research, because if we draw conclusions on theories that cannot be replicated, we diminish arguments and the strength of our science, making inferences on subjects that might not be there. File drawer effect and reproducibility can be said to be linked, as only publishing data that is significant, can be seen as increasing our rate of type I error. Even if our methods are rigorous, by skewing what we publish we are more likely to publish studies that committed some form of statistical error. Is there any way to combat this? Our proposed solution focuses on experimental design.

Experimental design is an essential component in ecological research, as in other disciplines. A general goal shared by many ecologists is to design experiments, whether conducted in natural environments or controlled laboratory settings. Such designs are aimed at minimizing biases and errors, such as type S (sign) and type M (magnitude) errors (Gelman and Carlin 2014). Moreover, the general

goals of experimental designs include increasing effect size and statistical power. Experimental designs also are tasked with discerning the nature of relationships between response and predictors, such as differentiating between linear and non-linear associations. Three common components of any experimental design are: (i) the number of treatments (levels or groups) within each factor and how they are distributed, (ii) the number of replicates per treatment, and (iii) the number of individual observations across treatments (Hurlbert, 1984). While there are other components that are critical to the design of experiments (e.g., random *versus* fixed factors, covariates to control for confounding variables), they are less relevant to the points we will be raising here. Ideally, the components of an experimental design will be decided deliberately and strategically. Failure to understand the implications of decisions about experimental design can reduce the ecologist's ability to properly design, analyse, and communicate their experimental results. These decisions may also have consequences for whether the results of an experiment are useful in systematic comparisons among studies (e.g., reproducibility, meta-analysis). Because experimental design is a core concept in research, ecologists should have a good understanding about some of the consequences underlying critical decisions underlying the design of experiments. While there has been calls for improved knowledge about how design decisions affect statistical inference (power and effect size) (Hurlbert 1984), but many of the issues remain unknown, misunderstood, or unclear (Petraitis 2005). Here we argue that ecologists, as a community, are overlooking some key design aspects that can have critical impacts on statistical power and effect size.

An important and well recognized issue that influences two critical statistical targets in an experiment, namely statistical power (i.e., probability of rejecting a null hypothesis that is false) and effect size, is the costs and trade-offs involved in allocating individual observations among the three above mentioned common design components. As resources are limited (e.g., time, space, funding, personnel) and ethical concerns need to be considered (e.g., number of animals involved), researchers must decide how to distribute the total number of individuals among replicates within treatments

versus among treatment levels. One central design question is how to balance these trade-offs while maximizing the ability to detect statistically significant effects (i.e., increase statistical power, decrease Type II error, maximize effect sizes) and understanding the relationship between response and predictors (treatment factors). While these decisions are crucial, authors often do not describe their decisional process in determining the number of treatments per factors and number of replicates per treatment within factors (Hulbert, 1984). Decisions underlying experimental design components are often made in non-judicious ways, leading to studies having low power and as recognized more recently, reproducibility (Baker 2016). While some studies undergo initial phases of power analysis in ecology, the average power of a study is generally low (Jennions and Møller, 2003). Although prior power analysis is in an important tool in identifying appropriate sample sizes, other critical decisions also can affect the power and effect sizes of an experiment. Most power analyses tend to focus on determining the smallest sample size needed to detect a significant effect given some estimate of the effect size estimated from pre-experimental data or based on other studies (Peres-Neto and Olden, 2000). Power analyses do not tend to be used, for example, on how to distribute treatment levels within a factor. Instead of simply focusing on increasing sample size to improve statistical power (significance), ecologists should be aware that for a given sample size, there are at least two other critical components that can affect the statistical power and effect sizes of a study, namely analytics (e.g., ANOVA versus regression, i.e how ANOVA puts more emphasis on between group variation, whereas regression focuses on variation between all points) and treatment dispersion (i.e., variation among levels of a quantitative factor). Note, however, that, in our experience with the literature, these are not the target of pre-experimental phases and can be useful in mitigating costly type II errors (Peterman, 1990).

Here, we focus on two main components that can increase the statistical power and effect sizes of experiments while not changing the total sample size (total number of observations) in an experiment: (i) the choice of analytics in which linear regression can, in many cases, replace ANOVA in

experiments involving quantitative factors. Linear regression is simply meant to express a regular regression used on replicated data as in observations within levels of a treatment (Zar 2022). We will demonstrate that regression containing replicated observations has greater statistical power and effect sizes when the relationship is linear (for non-linear relationships one could adapt Generalized Additive Model (GAM; Wood 2017) for the analysis of replicated data; and (ii) treatment dispersion (or sample density) in which we demonstrate how variance of a quantitative experimental factor affects the statistical power and effect size of both regression and ANOVA. Even though the first point has been discussed in the literature (though rarely, e.g., Cottingham et al. 2005), ecologists (among other researchers) are often unaware that the use of regression can increase statistical power and effect size. That said, most ecologists commonly resort to the use of ANOVA rather than (replicated) regression, even though the latter can be often directly applied on the same experimental data to increase statistical power. Cottingham et al. (2005) have well demonstrated that replicated regression can improve on ANOVA when both analytics are appropriate for the same data. However, they did not address the fact this improvement is a probabilistic expectation. That is, replicated regression is expected to be more powerful than ANOVA but there can be cases (data) for which the latter is more powerful than the former (see section 2 after Introduction for a discussion and recommendation). We argue that a critical component of experimental designs, namely the variance of predictors (i.e., dispersion of levels of quantitative factors or sample density) has been systematically overlooked by the research community as a key aspect affecting statistical performance (e.g., statistical power, effect size) and, in many ways, interpretability (e.g., directionality and non-linearity of effects). For instance, by having more treatments while reducing the number of replicates, we may gain a greater understanding of the relationship between response and predictor. The presentation and associated simulations here consider how the dispersion of levels affects the performance of ANOVA and replicated regression, and how these two analytics differ in this context. Model selection is integral to ecologists as this is how

they convey important information (Garaszegi et al. 2011). Although ANOVA and replicated regression are often considered similar models, experimenters can opt for the latter over the former in many situations using model selection and diagnostic procedures (e.g., linear versus non-linear fits). Some of the issues described here may be obvious to some but, in our experience, many ecologists still lack an in-depth understanding of the principles regarding how to optimize treatment and choice of analytics to improve statistical performance. We attempt to use language, simple fictional examples and simulations aimed at facilitating the communication of these issues. Our aim is to foster a deeper understanding of experimental design, assisting researchers in justifying their choices with stronger rationale. We also conducted a small meta-analysis on published data. An association would be expected given our prediction that treatment dispersion affects reported effect size, and any deviation would suggest that external factors may have significant contributions (e.g., previous experience with the study system and/or experimental settings).

It is important to recognize that maximizing treatment dispersion to increase reported effect size and power should be handled with nuance. The reason we must be wary of maximizing treatment dispersion, is that we must consider biological relevance when doing so, where biological relevance is defined as how any information deduced from an experiment is accurately portrayed in the natural system. Mathematically, the easiest way to increase treatment dispersion is to measure many values at extremes (Preacher et al. 2005). While this increases treatment dispersion significantly, the trade off is that we learn little about the system outside of our two extremes. In this sense, we do not necessarily have the information to make accurate inferences about the population outside of the extremes, nor if the population parameters in response to the treatment are truly linear or non-linear. Rather, it is important to pay attention to the amount of treatment dispersion in your experimental design when making inferences, and to think about the trade-offs between having high treatment dispersion, biological relevance, and feasibility of conducting an experiment in such a way. Having high effect size

and power in an experiment where the relevancy of the experiment is in question can contribute to poor reproducibility, as the extrapolation knowledge learned from extremes can be useful but not without risks such as data dichotomization (Preacher et al. 2005). Rather we must conclude that treatment dispersion should be considered as a parameter that influences the effect size and power, and that trade offs in design should be considered with treatment dispersion in mind.

How many replicates where? Understanding the trade-off between number of replicates and number of levels within factors

While one obvious solution to improve experimental designs is to increase total sample size by increasing the number of treatment levels and number of replicates per level, one must find a cut-off point where the increase is most efficient to avoid funding limitations and ethical issues. For simplicity, we describe a replicate as an individual observation (e.g., individual fish, one single plant) rather than a group of individuals within a block (e.g., individuals within a tank). However, often in ecology, groups of individuals within a block (e.g., tank) within a treatment represent a replicate. In this case, a treatment level can have multiple replicates (e.g., multiple tanks with multiple individuals each for a temperature level). In this case, a mixed model is appropriate and would treat replicates (tanks) as a random effect. That said, the issues described here apply regardless of how replicates are defined. The most obvious decision when designing an experiment is likely setting the number of levels for any given factor (treatment, i.e., controlled predictor) and the number of replicates for each level. But even then, challenges exist. Quoting Sefer et al. (2016), “Given limited budget, should we profile more repeat experiments or more time points?”. Assuming a trade-off between number of replicates and number of levels (often referred as sampling density) exists, one may be tempted to consider a smaller number of levels for a given treatment to allow greater number of replicates per level. And because manipulating

the number of levels in a factor may be more costly and/or challenging (e.g., keep multiple aquaria at multiple temperature levels) than the costs in handling multiple replicates, one may be tempted to increase the number of replicates per level instead of increasing the number of levels. While increasing sample sizes will increase statistical power, it may not be the most efficient way to manage limited resources (Sullivan et al. 2012) and ethical considerations (e.g., animals used in experimentation). Perhaps a critical point to consider here is that by establishing designs that lead to appropriate statistical power and effect size, one reduces the future costs of validation and reproducibility that may be necessary to assess previously non-rejected hypotheses (e.g., negative results). This may be used as an argument towards increasing total sample sizes (e.g., individuals used in a single experiment). As an example, consider an experimenter who wants to measure a response variable over a wide range of temperature levels. For practical reasons, they can only consider a limited number of intervals (levels), so they increase the sample size per level (number of replicates) hoping for an increase in the statistical power of the experiment. While researchers might consider the decisions on the number replicates (replicates within the same level of a factor) as the lowest hanging fruit component in determining effect size and statistical power, other components can have greater influence as we demonstrate here. Nonetheless, a simple demonstration (next) of this issue seems relevant for our purposes of improving knowledge about the issues underlying the design of experiments.

Assume an experiment where the total number of replicates is 36 observations (e.g., individual fish) and that the effect size is 0.26 for a given factor of interest (e.g., effects of factor temperature on response variable fish growth). A commonly used effect size for ANOVA (Cohen's *f*) is $f=\sqrt{\eta^2/(1-\eta^2)}$ where η^2 is the sum-of-squares of treatment (factor) and $1-\eta^2$ is then the sum-of-squares of residuals. If that same effect size *f* is the outcome based on an experiment with 2 temperature levels (say low and high) with 18 replicates (individual fish) each, then the power of the ANOVA (i.e., probability of rejecting the false null hypothesis) is 0.60. However, if the experiment were to be based on 6 temperature levels

with 6 replicates (individual fish) each, then the power is much higher at 0.86. Hence, as the number of levels of a factor increases while keeping constant the total number of replicates (assuming the same desired effect size), the probability of achieving a significant effect increases (i.e., statistical power). That said, manipulating the number of levels in a factor may be more costly, hence the trade-off in choosing between number of replicates and number of levels of a factor given a total number of replicates (i.e., across all factors and levels) in an experiment. As such, the number of levels and replicates per level has important consequences to statistical power and generating information about differences sources of variation (see Petraitis 1998, Blainey et al. 2014) as we discuss in the next sections.

Variance of quantitative predictors (and ANOVA factors) increase statistical power and effect size:

ANOVA *versus* regression – a small distinction with big potential gains

The robustness of regression over ANOVA when both analytics can be used on the same data (see discussion on model selection) is true even for small number of treatments. In experiments with as few as 3 levels (groups or treatments) per factor (say an experiment using three temperature levels), the use of a regression over an ANOVA would increase power and effect size (Cottingham et al. 2005). The bottom line is that even though designs are specifically set up having an ANOVA model in mind, replicated regression may increase statistical power and effect size when dealing with continuous (and ordinal) data. As noted earlier, the key to understand the power and effect size differences between the two models is that ANOVA uses differences of variance between and within levels (groups/treatments), whereas regression uses the slope of the response across multiple levels. ANOVA is then less sensitive to large differences within measurements of predictor variables. Since regression is sensitive to slope variation and ANOVA to level differences, when the assumption of linearity is met, regression is more sensitive to these differences (Cottingham et al. 2005). Therefore, as treatment (level) dispersion increases, the difference in statistical power between regression and ANOVA also increases. However,

regression is more powerful only when the assumption of linearity is met. ANOVA, though a linear model, it is based on unordered predictors and not affected by non-linearity. ANOVA is in a way spatially insensitive, whereas regression is spatially sensitive.

ANOVA is popular despite the obvious power differences because it performs well for non-linear relationships between treatments (unordered) and response. Let us consider here three small fictional examples. Figure 1A represents a case in which both ANOVA and regression are appropriate, but regression (p=0.020) has greater statistical power over ANOVA (p=0.238). Figure 1B represents a case where both ANOVA and regression are appropriate, but ANOVA (p=0.025) has greater statistical power over regression (p=0.478). Note that, regression has greater statistical power than ANOVA (see our simulations below). That said, this is a probabilistic expectation, and even when the true trend is linear, an ANOVA may still have a lower p-value than a linear regression in a specific replication (Figure 1B). One could make the case here that is worth considering both approaches and pick the one with the largest effect size and smaller p-value (we discuss this point in greater detail later). It is critical to note that Cottingham et al. (2005), who made the point that linear regression is more powerful than ANOVA, did not address the issue of power expectation. This may have led some ecologists to believe that regression analysis is always more powerful than ANOVA. And finally, Figure 1C represents a case where ANOVA (p=0.001) is the appropriate analytics over regression (p=0.138) due to the non-linear nature of the data. Finally, we fit a Generalized Additive Model (GAM) to the data in Figure 1C (non-linear) to demonstrate the potential of a non-linear replicated regression as analytics (p=0.0001). Although we will not be covering non-linear regression here, we feel that is critical to mention that they can improve the ability of researchers to seek further ways to improve their knowledge about experimental design and the analytics involved; as well as properly describe response surfaces. In fact, given the robust penalization procedure in GAM, one could always just fit a GAM model to replicated data instead of contrasting whether a linear model would fit best than a non-linear one. This is because if the nature of

the model is linear, GAM will be penalized to fit a simple linear regression model (Guisan et al. 2003). Cottingham et al. (2005) made the point that for non-linear replicated relationships, ANOVA should be preferred over regression. However, they did not consider the case of non-linear models (e.g., GAM) which we bring to attention. Note also that although non-linear approaches to regression are expected to be more powerful, ANOVA may lead to greater effect sizes and smaller p-values (i.e., greater power). As such, one can conduct ANOVA and GAM, retaining the one that led to the greatest power. Another advantage of GAM not shared by ANOVA is that it could allow for greater replicability in the sense that the response curves across equivalent experiments can be more easily contrasted using GAMs than ANOVA.

Analysis of variance (ANOVA) and linear regression are likely the most common experimental designs used in biology (Konietschke et al. 2012) and though they are similar mathematically, they are not the same inferentially (i.e., under statistical hypothesis testing) as regression has greater degrees of freedom in contrast to ANOVA. As such, regression is expected to have great power than ANOVA. This key difference can have drastic effects on statistical power. While regression is a general linear model that uses the slope to estimate differences between data points, ANOVA is a general linear model that estimates variation within and between groups (e.g., experimental levels). Most researchers apply ANOVAs by measuring a response variable in relation to an independent variable (factor) that is divided into levels (i.e., that are organized into categories, e.g., low, intermediate, and high temperature). Note that ANOVAs are also applied to regressions involving continuous independent variables only (Ritcher, 2006), which will become important for our arguments later as many experiments involve transforming continuous variables into categorical factors. In fact, although regression and ANOVAs involve the same algebraic approaches, they can differ in their inferential capabilities due to experimental design (see Cottingham et al. 2005 for an in-depth discussion). The use of ANOVA (and t-tests) to tackle continuous

variation (response) across discrete predictors (groups or treatments) is so pervasive that they are amongst the first inferential tools we learn in introductory statistical courses. Researchers are often accustomed to categorizing continuous predictors (factors) rather than using them in their original continuous form as in a regression analysis. The former (i.e., discretization) is often seen (though inaccurately) as a cost saving measure. Consider that a researcher wants to measure fish metabolism as a response to temperature. Instead of measuring and making tanks to hold constant temperature across a very wide gradient of temperatures (akin to what many think as a requisite for a regression), it would seem more cost efficient to measure two or three different temperatures only. While is known that regression can lead to greater statistical power for the same data in contrast to ANOVA (Cottingham et al. 2005), it may be rather costly or unfeasible to run experiments over many temperature treatments where the differences between ANOVA and (replicated regression) are more noticeable. However even in cases where we consider running experiments over limited temperature ranges, a (replicated) regression leads to increase effect sizes and smaller p-values (i.e., greater statistical power).

Treatment dispersion - variance of quantitative predictors increase statistical power and effect size of ANOVA and regression.

Perhaps the most common way in which researchers think about and conduct power analyses for ANOVA is: given an observed effect (i.e., mean variability among treatments) and the error sum of squares (i.e., mean variability within treatments), how does statistical power change as a function of increasing number of replicates (observations) within levels (treatment/group). Mean variability among and within treatments are then used to estimate power and effect size. Although less commonly available in software, effect size can be also estimated (leading to identical results) by providing the values of the maximum difference between means and standard deviation for samples (Cohen 1988; e.g., Minitab). These maximum differences can be used to vary treatment dispersion as well. Consequently, most researchers do not consider manipulating treatment dispersion even though it is

possible to do so. Next, we discuss the consequences of treatment dispersion to ANOVA and replicated regression.

The ANOVA effect size has connection with the well-known coefficient of determination (R^2; Cohen 1988; see also Cottingham et al. 2005). Cohen (1988; also based on some previous work referenced there) demonstrated that the effect size for ANOVA can be expressed as $f = R^2 / (1 - R^2)$. Because R^2 is mostly used in regressions, many empiricists are unaware of its connection with ANOVA, even though the R software for statistical computing can easily output this value and other software can produce the necessary quantities for calculate R^2 for an ANOVA. This is important because R^2 makes the connection between ANOVA and regression more direct. There are different ways to express the same F-statistic; one that is useful here to make the connection with R^2 is:

$$F = \frac{(SS_T - SS_E)/(p-1)}{SS_E/(n-p)}$$

where SS_T is total sum of squares, SS_E is the residual sum of squares, *p* is the number of factors (ANOVA terminology) or predictors (regression terminology). Based on the F statistic, R^2 then equals:

$$R^2 = 1 - \frac{1}{1 + F \cdot \frac{p-1}{n-p}}$$

The connections between F and R^2 is critical for us to demonstrate that the statistical power of ANOVA is influenced not only by degrees of freedom (related to *n* and *p*) but also by the treatment dispersion (closely related to regression) or among group levels (closely related to ANOVA). Treatment dispersion represents the variance among treatment values involving continuous predictors. The effects on effect size and power of treatment dispersion on ANOVA and regression do not appear to be a well-known fact among empiricists and has critical consequences for the design of experiments. Instead of using an algebraic demonstration of how variance of predictors influences R^2 (and necessarily effect

sizes in ANOVA and regression), we will use simulations instead (see Methods and Results sections). The main goal of our simulations in this study is to provide an intuitive, rather than a mathematical, demonstration that the distribution of levels (groups/treatments) within factors (treatment dispersion) can have potential impacts on statistical power and effect size.

Reducing the treatment dispersion and/or variance of predictor values reduces both statistical power and effect size. Figure 2 illustrates a regression whereby reducing the variance of the predictor drastically reduces both statistical power and effect size (see legend of Figure 2 for summaries). It represents two data sets that share about the same slope but differ in variance of predictors. The data with larger predictor variance (Figure 2) has greater power and greater effect size in contrast to the data with the smaller predictor variance. This extreme example is useful to demonstrate that the same slope can lead to very different inferential conclusions. However, less extreme changes can cause substantial shifts in statistical power and effect size as demonstrated in Figure 3. This figure represents two experiments that have the same slope but differ in treatment (level) dispersion. The experiment with larger treatment dispersion (Figure 3, experiment 1) has greater power (smaller p-values) and greater effect size (see Table 1). This contrived example is useful to demonstrate that the same slope can lead to very different statistical conclusions, i.e., the experiment with the larger dispersion was significant whereas the one with the smaller dispersion was not. Since we selected the slope beforehand, we know that both experiments exhibit the same relationship, and therefore one could have assumed that the results would be more comparable. Note that the fictional data in Figure 3 had different variances and different ranges. If we control for range of treatment levels, we still observe similar results (see Figure 4 and Table 2). Both cases (Figures 3 and 4) used a selected fictional data set that illustrates well (in an exaggerated form) the effects of treatment dispersion. However, there are certainly many situations in which effect size varies between different treatment dispersion values but the conclusion underlying statistical significance is unchanged (i.e., they achieved the same statistical power given a certain alpha

level). Our simulations later are designed to demonstrate how treatment dispersion affects effect size and statistical power over a larger range of situations to quantify the degree of these differences.

Overall study design

As previously mentioned, using ANOVA over a regression will (in general) result in lower power and effect size for linear models (Plonsky and Oswald, 2017), and the differences regarding these components are expected to increase between analytics with treatment dispersion. In our simulation study, we conducted ANOVAs and regressions on simulated datasets and experimental designs where the F-value, effect sizes and p-values of each test are calculated and stored. Aspects that we control in each design are the following: slope, treatment dispersion, number of replicates per level, and number of levels per treatment. Slope was set to a fixed value that varies per simulation and was always greater than 0 to meet the assumption that the null hypothesis was false and therefore allowing estimating power rates. In this way, the stronger the slope the stronger the effect size of the treatment. Effect size and minimum sample size have long been known to increase power and provide adequate design (Green, 1991). Number of replicates and levels per treatment were also manipulated. Using this simulation design, we conducted a power analysis by comparing the percentage of p-values that were significant for an assumed alpha level of 0.05 over the total number of tests for a given scenario (here set to 1000).

Parameters such as variance of treatment, replicates per level, and levels per treatment are all determined by the researcher and, as such, we hope that this study servers as a guide on how to improve experimental design. We will test the predictions that: maximizing all three parameters will lead to the greatest increases in power; and high treatment dispersion increases statistical power. Such a result would suggest that even using smaller sample sizes, a certain acceptable power (e.g., 80%) can be achieved depending on the treatment dispersion. This knowledge can help provide economic

solutions to resource limited research programs, thus increasing research accessibility. By reducing sample size and optimizing treatment dispersal, researchers can also minimize animal utilization, thus improving experimentation ethics.

One important question is whether researchers possess some innate intuition about the fact that increasing dispersion treatment leads to increased effect sizes. We assume that this knowledge would be well known or even intuitive, given that none of these studies (as is the case of most studies using experimental designs in our experience with the literature) explain the decisions underlying the choice of numerical values for treatment (i.e., treatment dispersion). Behavioural studies commonly underly decisions concerning number of observations given ethical concerns implicated in research using animals. Live animals such as fish can be expensive to manipulate while keeping large number of individuals in a healthy environment. We predicted the meta-analysis to show the same results as the simulation, in that experiments with large treatment dispersion should also lead to greater effect sizes and likely to increased statistical power. That said, other factors may affect our expectations, including the fact that two experiments could achieve the same effect size for different reasons (e.g., one increasing sample sizes and the other dispersion treatment).

Methods

All simulations were performed in the R environment (R Core Team, 2017; version 4.1.2). All packages and functions are referred here in italics. Here we considered only the case of linear relationships between predictor (treatment levels) and response. This served to compare regression and ANOVA more directly and to manipulate the strength of the relationship more easily via a single regression slope for any given scenario. As discussed earlier, however, one could use non-linear approaches to regression for analysing experimental data involving continuous predictors; but, for simplicity, we did not consider the simulation of non-linear relationships. Simulation results were

reported in two ways, one where we assessed the relationship between effect size and standard deviation, and another where we conducted a power analysis. The effect size metric was calculated as partial omega squared (see Lakens, 2017) and power analysis was conducted by measuring the proportion of simulated replicates for the same scenario (i.e., combination of parameters) that achieved statistical significance under a significance level (alpha) of 0.05. ANOVAs were conducted using the *aov* function and regression using the *lm* function, both part of R Base. Manipulated simulation parameters included the number of treatment levels, the number of replicated (individual observations) per level, the slope between predictor and response, and variance of the independent variable (treatment dispersion). The values for each parameter were as follows. Treatment variance (dispersion) assumed 1, 3, 5 ,7, and 9; slopes were set as either 0.05, 0.1 or 0.2 Number of levels were set either as 5, 10 or 15 and number of replicates per level depended on the total number of individuals (either 30 or 60) considered in each "experiment". For example, if 5 levels were considered, then either 6 replicates were generated when 30 observations in total were considered; or 12 replicates per level when 60 observations in total were considered. Keeping the total number of observations at these two levels (though more levels could have been considered) allowed considering the case that the resources (costs) invested in terms of total number of individuals in each "experiment" were kept constant. In total, 72 scenarios were considered (4 treatment dispersions x 3 slopes x 3 number of treatment levels x 2 total number of observations) and for each scenario, 1000 simulation replications were performed. Data generation and analyses for each simulation scenario (i.e., unique combinations of the four parameters) were set as follows:

1) Generate values for each treatment level (predictor). We started by generating an initial set of means using normally distributed random values (assuming zero mean) according to the desired number of treatment levels and desired variance (dispersion treatment). Note that this mean does not affect statistical power or effect sizes. The range of these means were then divided into equally distributed

intervals using the same number of desired treatment levels. We then picked the lower bound value of each interval and used those as the treatment values. Note that, for simplicity, only one factor was considered in all simulations. If the centre or the upper bound value were used, the same results would have been achieved.

2) Generate the expected response mean values according to the bound values for the treatment levels (step 1) by multiplying the latter by the desired slope.

3) Finally, response values were simulated by generating normally distributed random values (assuming unity variance) around the expected response mean values (step 2) according to the desired number of replicates (i.e., sample size). The number of replicates was also manipulated to vary among scenarios (see below). The expected variance within treatment was set to one (unity variance) across all scenarios (i.e., homoscedasticity). Obviously, the variance within treatments can affect effect size and statistical power. However, given that this variance cannot be set by the researcher we simply assumed a common value so that results can be easily contrasted across all scenarios.

We used partial omega squared as a metric of effect size (see Lakens, 2013 for calculations). The average effect sizes were subsequently calculated for various iterations of replicates per treatment, number of treatments, variance of treatment, and slope.

Meta-analysis – do researchers manipulate treatment dispersion to improve inference?

The goal of the meta-analysis is to assess whether researchers may (instinctively or perhaps deliberately) manipulate treatment dispersion to increase effect size and/or statistical power. Again, because the information on how precisely treatment levels (treatment dispersal) were selected is rarely described, we felt compelled to assess whether there may some inherent knowledge about how dispersion treatment affects effect size and statistical power. Although our meta-analysis is restricted in terms of number of studies, it serves as a potential source of discussion for this paper and/or for future

studies. We focused on fish behavioural studies as they commonly involve categorizing (discretizing) continuous variables. There are many reasons why fish behavioural ecologists are inclined to categorize variables, but one of them is that continuous variables such as temperature are easier to control and compare across aquarium environments. Because the goal of our meta-analysis was for demonstration purposes, papers were searched using Google Scholar, and search terms used to target papers had two components: the species of fish being studied and the year of publication. For our meta-analysis, we targeted three common groups of fish that are often used in experiments: guppies, cichlids, and salmonids. To target these groups, the terms included "Guppies and Behaviour", "Cichlids and Behaviour" and "Salmon and Behaviour". Years covered ranged from 2017 to 2020, and specific years were targeted by setting the search only to these years. Depending on success of search, anywhere from 5 – 10 papers from each category were identified as candidates to be included in the meta-analysis. Again, the goal of our meta-analysis was not to be systematic but rather provide (or not) evidence whether researchers may be manipulating treatment dispersion.

Inclusion in the meta-analysis (see Table S1 in supplement material for list of studies) had to use some form of linear model that reported test statistics that were either F values or could be converted to F-values (t-statistics can be converted to F easily). Only papers that used an ANOVA were considered for the analysis, and any papers that used regression were excluded. Although rarely used in experiments, researchers using replicated regression might be more aware of how treatment dispersion affects inferential outcomes. Furthermore, the continuous explanatory variable must have been discretized into categories so that we could calculate the standard deviation of the explanatory variable. In addition to F-values, we also recorded p-values, denominator degrees of freedom, numerator degrees of freedom, levels, number of factors in the model, species, dependent variable, standard deviation of the explanatory variable, type of statistical test, standard deviation of the explanatory variable, and

number of replicates per treatment. We used partial omega-squared to calculate effect size as it uses F-values and degrees of freedom in its calculations (which are readily available from the studies).

A Linear Mixed Model (LMM) was applied to explore the relationship between effect size and ranked standard deviation within the meta-analysis. Publication was considered as a random factor as multiple values were often recorded from the same study. To meet the assumption of normality of residuals, effect size underwent ranked transformation. Normality and heteroscedasticity were assessed utilizing Q-Q plots and Shapiro wilk test, respectively. Simulations and analyses were performed in R and packages include *ggplot2, dmetar, glmmm, meta, metafor, tidyverse, and jtools*.

Simulation results

The results of the simulation were as expected and served to demonstrate the issues we wanted to raise. Increasing dispersion (variance) of treatments led to power increases in both ANOVA and regression for all cases (Figure 5). Note again that the total number of observations were kept constant for a given number of treatments (Figure 5; either total of 30 or 60 observations). For a given total sample size (i.e., either 30 or 60 observations), increasing the number of treatments (i.e., treatment levels) led to power increases in regression, but less pronounced increases in ANOVA (Figure 5). This is to be expected, given that the variance in our simulations increase slightly with the number of levels in a treatment (see Methods). If the treatment levels would have been scaled to unit variance, then the difference in power would change slightly; regression would still have greater power as number of treatments increase and ANOVA would decrease in power as number of treatments increase. This is because ANOVA, unlike regression, penalizes (via degrees of freedom) the number of treatment levels. Note, however, that our interest was mostly to contrast how slope and variance affected regression and ANOVA power and not so much the number of treatments. That said, it is likely that in real experiments,

the variance among treatments may increase with the number of treatments as controlling levels to have the same variance would be difficult.

Power increase with number of replicates (i.e., sample size per treatment) can be considered intuitive in practice. Increasing sample size has long been known to increase statistical power (McCrum-Garder, 2010), reflecting the fact that a larger sample should reduce sampling error. If a sample is too small, the chance that the sample misrepresents the makeup of the population increases drastically. The only drawback to increased sample size is resources (and ethics in case of studies involving animals), as it has no effects on type I error, unlike other ways of increasing power (e.g., increase dispersal treatment; Wang et al., 2017). The effect of slope (which modulates the strength of the effect) on power may also seem obvious to the informed observer. Essentially, the stronger the effect the more likely it is to detect it in both ANOVA and regression analyses (Figures 5 and 6). Since we are comparing the slope magnitude against zero, the larger the difference, the lower the chance a test would incur a type II error (i.e., reduce power). However, this has ramifications for studies where small differences in treatment level ("cause") can have big effects (i.e., induce changes in the response variable) (Wissuwa, 2003) and, as such, these studies need to increase power using ways other than increasing sample size (Hansen and Collins, 2004). A question that may come to mind is why ANOVA, which does not take into consideration the slope, has greater power because of an increase in slope in our simulations (Figure 6)? This is because as slope increases, the expected mean values across treatment levels. Our simulation demonstrates this point well by showing that the increase in power in relation to slope variation for ANOVA is far less than the power increase of regression (Figure 6 and 7 contrasting ANOVA and regression). In this case, while regression is directly influenced by slope, ANOVA is only indirectly influenced by it, offering relatively smaller increases in power.

As for statistical power, effect size, a metric generally used to describe the magnitude of experimental results, is also influenced by design choices (Figure 7). Researchers interested in meta-

analysis may find that studies with the greatest effect sizes are biased towards those that considered large dispersion treatment levels. Additionally, partial omega-squared increased more noticeably for regression then ANOVA given the same slopes (Figure 7). As such, combining results from regression and ANOVA in the same meta-analysis based on their effect sizes may prove challenging. This was one reason for why we concentrated on ANOVA rather than regression in our meta-analysis, particularly given that the former is the more common analytic in experimental designs.

Meta-analysis results

Contrary to our predictions, the meta-analysis showed that there was no statistically significant relationship ($t = -1.27$, $p = 0.23$) between effect size and treatment dispersion (variance of treatment levels) in ANOVA based fish behavioural studies (Figure 8 and Table 3). Most of the variation can be explained by published paper (random factor), as a paper reporting multiple effect sizes would often report similar effect sizes (Pseudo-R^2 of random effects = 0.57; Table 3). As such, the likelihood of a paper to report similar effect sizes would then seem high, even though each effect size may relate to a different treatment effect.

Discussion

The simulations largely met our predictions, as power and effect size increased with treatment dispersion and slope. The magnitude of the difference in statistical power between ANOVA and replicated regression increased with both slope and variance of treatment levels (treatment dispersion). While researchers generally have little or no control over the magnitude, shape, and direction of experimental outcomes, they can exert strong control over treatment dispersion. Therefore, our simulations imply that not only do we have an alternate strategy to increase statistical power without needing to increase sample size, but we also have a way to directly increase effect size as well (though the non-linear relationship between p-value and effect size is well established). The ability of

manipulating dispersion treatment to increase effect size and power brings some important considerations to mind. For instance, one may increase variance of treatment levels by making some treatment levels too similar and others different as illustrated in Figure 3 and Figure 4. Although this may increase the power of ANOVA and regression, it may not necessarily assist with post-hoc comparisons (e.g., two by two contrasts between means, which would result exactly on the same values for both ANOVA and regression).

Our simulations revealed critical differences in power and effect size between ANOVA and regression. Although this was expected based on statistical knowledge, most researchers are probably unaware of their inferential properties and resulting differences. As Cottingham et al. (2005) pointed out, ANOVA should be preferred when the relationship between treatments and response are non-linear as contrasts are used in an unordered fashion. However, it is unlikely that researchers consistently use ANOVAs uniquely because of the linearity assumption of regressions. Instead, researchers should be compelled to first check the assumption of linearity and then chose the more powerful analytics for their data (e.g., replicated linear versus replicated non-linear regression; Figure 1C). Unless non-linearity leads to increased type I error when using a linear model, one could argue that using regression is the better model if it results in greater effect sizes and smaller p-values. There are obviously other concerns than just the loss of power due to non-linearity when using ANOVA over regression. Considering whether data are continuous or categorical should be a clear consideration. If data are truly categorical (i.e., not discretized for convenience to apply ANOVA), then ANOVA would certainly be the correct analytic over replicated regression. However, as it is often the case, one is often tempted to discretize treatment levels for ANOVA convenience. Linearity does arise as an important issue, even though one can consider non-linear replicated regression approaches (Figure 1C). More importantly, one should recognize the pros and cons of ANOVA and regression and make decisions accordingly.

meta-analysismeta-analysis It is important to reiterate here what assumptions our simulations made and how relevant they are to natural systems. Ours simulations assume linearity which may be seen as a restrictive because one could make the point that most systems in biology are likely described as, at least, partially non-linear. As such, experiments with large treatment dispersions have the benefit of increasing information about the nature of the relationship between cause (treatment) and effect (response) if an adequate number of levels are used. Even though variance can be drastically increased by considering a few levels, the nature of the relationships may be uncovered. For instance, in the extreme case in which only two levels are considered, the relationship will always be linear. Two components play a role in increasing treatment dispersion, namely range and distribution of levels (Figures 3 and 4). Distributing treatment levels uniformly within the same range (Figure 3) may decrease statistical power though, conversely, it may improve knowledge about the relationship between response and predictor (response surface). These are critical aspects to be considered. As many researchers consider statistical power important to determine in pre-experimental phases, we argue that the distribution of levels accounting for both dispersion and uncovering the relationship between predictor and response should be also relevant in pre-experimental phases; and critical in experiments set without pre-experimental phases.

In this study, we considered the contrast between ANOVA and regression, particularly because the choice between the two do not seem related to lack of knowledge of the two analytics, but rather how they contrast; a point that we hope that was made clear in our simulations. However, it is important to mention that ANOVA and regression are not the only options in terms of analytics for experiments. Generalized linear models (Nelder and Wedderburn, 1972) may be preferred depending on the nature of the response variable (e.g., ratios, abundances) and robust approaches to linear and non-linear regressions (Lim et al. 2013) to deal with outliers, among other issues. Much like how regression has higher power over ANOVA in linear responses, most forms of non-linear modelling should

be more powerful than ANOVA when applied to non-linear response surfaces. As such, making informed decisions about data distributions and choosing the appropriate model is more beneficial than using ANOVA as universal analytics for experimental data. However, prior knowledge about data outcomes is likely highly variable among researchers, and this can affect the design of an experiment.

Contrary to our expectations, our meta-analysis showed that effect size seems not vary as a function of treatment dispersion; though not significant, the relationship was only slightly positive, i.e., as treatment dispersion increases, the effect size increases by a negligible amount. One issue to consider is how researchers are knowledgeable about their study systems, or how much prior research has already been performed and assist in their decisions prior to setting an experiment. If a system is extremely well known, particularly when experimental decisions are well reported and adjudicated, researchers will likely have a good understanding on how to improve designs to achieve statistical power and large effect sizes. For example, it would make no sense to inject a drug past its lethal dose, where a lower dose may elicit the same response. Increasing the dose would lead to a lower effect size if the effect did not increase by the same margin. This is called a ceiling effect (Wang et al., 2008), a statistical principle that stipulates that past a certain point, whatever effect you detect will not change even if you measure beyond that point. And strong ceiling effects reduce statistical power and effect size. So, for a system in which behavioural responses are well understood, it makes sense to avoid ceiling effects as much as possible by using that knowledge to set treatment dispersion. Ceiling effects should be accounted for (Walsh et al. 1994) and the effects of a treatment should then focus on lower levels rather than higher ones even though effect size may be smaller. However, in systems where prior knowledge is sparse or unavailable, researchers may be inclined to increase treatment dispersion blindly. One then wonders if the specifics of ceiling effects to a particular system are well known by the researchers in that field. Not considering ceiling effect in a particular experimental system judiciously

may explain, at least in part, why we did not detect a relationship between effect size and treatment dispersion.

The results of the Meta-Analysis may in fact suggest that there might be some form of file drawer effect influencing results. We would expect that at low dispersion we would get low effect size, but unexpectedly the effect size is similar to results found at high dispersion. While not conclusive, this can potentially indicate that studies with low effect size at low dispersion might have lower publication rates, influencing the nature of results. In fact, the results suggest that most studies that are published have high effect size, regardless of their dispersion values. This is directly in contrast to the simulation, which showed that as dispersion increases, so too does reported effect size. Consequently, reproducibility of these experiments could potentially be lower, as the reported effect size average is higher than expected. However, there are many reasons that would suggest that reproducibility is not necessarily affected. For example, it could be that the study system is well known, and lower dispersion should consequently yield expected high effect sizes, or rather the power of the experiment (albeit not as high as a study with high dispersion) is still quite high, and since the true effect size is large, that is what ends up correctly being reported.

What if we consider these results from a point of view in the worse case scenario; that file drawer effects are quite rampant in published studies. That reality would force us to carefully reconsider the review process, and what papers we consider “publishable”. Should we publish studies without much narrative for the sake of increasing our absolute knowledge on the systems published. One might argue that by doing so we lower our standards for what is published, and we may lose important findings to a diluted, albeit more accurate pool of papers. However, the decline in academic standards should not be a problem, as the papers published would still need to adhere to the review process’ criteria of scientific robustness, i.e. they must have a rigorous scientific method and a well thought out experimental design. How about a diluted pool of papers? We would argue that the concern of dilution

is a concern that published papers would not have interesting narratives. Unfortunately, (and depending on individual tastes) to reaffirm a known hypothesis can be considered less interesting than supporting a new one, and this may have impacts on the way journals engage their readers. This is shown in the way journals use impact factors as an important metric in article quality, which favors groundbreaking findings, over replication of known phenomena. Luckily as mentioned, this would be a worst-case scenario, and we am not claiming many papers today suffer from a file drawer effect based on my meta-analysis results. Rather we must consider file drawer effect as a potential explanation that surely has some contribution to these results, and to be aware of how our systems that are currently in place favor file drawer effects.

Another potential explanation underlying our meta-analysis result is the fact that study systems may be often non-linear, particularly if they have a ceiling effect. Moreover, reducing treatment dispersion can lead to wrong inference based on linear relationships for non-linear phenomena. Again, an obvious example is a study system with only two levels. No matter how great the treatment dispersion, the researcher will always find a linear relationship. As a result, a non-linear system is expected to have a lower effect size (on average) even under large treatment dispersion (Ganzach, 1998) (especially if the relationship oscillates or plateaus). Increasing the number of levels does better at capturing the non-linearity of the system, but high treatment dispersion may not lead necessarily to increase in power and effect size due to non-linearity. Another factor to consider is oscillation (low-high-low-high across treatments, i.e., non-linear systems). In this case, low treatment dispersion may lead to increased statistical power and effect size (though lower knowledge about how the treatment affects the response). Capturing non-linearity in experimental ecology is challenging and it has been shown that non-replicated sampling across gradient designs (linear or non-linear) can have higher success (Kreyling et al. 2018; in their simulations, success represented increases in R^2). Non-replicated designs can increase both treatment dispersion and ability to better capture response surfaces. Taken together, the

lack of relationship between treatment dispersion and effect size found here may be due to variation in the way response surfaces vary among studies. This is because effect sizes estimated from standard ANOVA and non-linear replicated regression for the same data would differ. A more systematic way of describing response surfaces (e.g., non-linear replicated regression) would be also beneficial within individual studies and critical for meta-analysis.

Conclusion

We have hopefully demonstrated the importance of increasing critical thinking about the design of experiments. If the response surface is linear, regression is expected (in average) to yields higher effect size and power than ANOVA and should, therefore, be considered when treatments are continuous. However, as Figure 1B well demonstrated, there can be cases in which an ANOVA analysis results in a lower p-value even when the assumptions of simple linear regression are met. Therefore, it is worth using both and picking the one that leads to the highest performance. Our meta-analysis allowed reflecting on important issues among studies. The first one being how to contrast effect size among studies that considered very different treatment dispersions. The second is how consider variation in response surfaces while describing experimental results and contrast of effect size, particularly when there can be a complex link between the shape of the response surface and variation in treatment dispersion. We hope that our study and narrative motivate the notion that better understanding different components affecting experimental designs and their outcomes is critical; and that often the same data can lead to different inferential outcomes (e.g., statistical significance and effect size) simply based on analytical choices rather than only design.

References

Asifa, K. P., & Chitra, K. C. (2018). Evaluation of acute toxicity of octylphenol in the cichlid fish , *Pseudetroplus maculatus* (Bloch , 1795). *International Journal of Applied Research*, *4*, 197–203.

Baker, M., & Penny, D. (2016). Is there a reproducibility crisis? *Nature*, *533*, 452–454.

Baerends, G. P., & Baerends-Van Roon, J. M. (1950). An introduction to the study of the ethology of cichlid fishes. *Behaviour*, *77*, 199–199.

Blainey, P., Krzywinski, M., & Altman, N. (2014). Quality if often more important than quantity. *Nature Methods*, *11,* 879-880.

Brandão, M. L., Colognesi, G., Bolognesi, M. C., Costa-Ferreira, R. S., Carvalho, T. B., & Gonçalves-de-Freitas, E. (2018). Water temperature affects aggressive interactions in a Neotropical cichlid fish. *Neotropical Ichthyology*, *16*, 1–8.

Buckless, F. A, & Ravenscroft, S. P. (1990). Contrast Coding: A Refinement of ANOVA in Behavioral Analysis. *The Accounting Review*, *65*, 933.

Butler, J. M., Whitlow, S. M., Roberts, D. A., & Maruska, K. P. (2018). Neural and behavioural correlates of repeated social defeat. *Scientific Reports*, *8*, 1–13.

Cattelan, S., Lucon-Xiccato, T., Pilastro, A., & Griggio, M. (2017). Is the mirror test a valid measure of fish sociability? *Animal Behaviour*, *127*, 109–116.

Cogliati, K. M., Unrein, J. R., Stewart, H. A., Schreck, C. B., & Noakes, D. L. G. (2018). Egg size and emergence timing affect morphology and behavior in juvenile Chinook Salmon, *Oncorhynchus tshawytscha*. *Ecology and Evolution*, *8*, 778–789.

Cohen, J. (1988) Statistical Power Analysis for the Behavioural Sciences, 2nd ed. Erlbaum, Hillsdale, NJ.

Cottingham, K. L. (2005). Knowing when to draw the line: designing more informative ecological experiments, *Ecological Society of America*, *3*, 145-152.

De Serrano, A. R., Daniel, M. J., & Rodd, F. H. (2021). Experimentally altered male mating behaviour affects offspring exploratory behaviour via nongenetic paternal effects. *Behavioural Brain Research*, *401*, 113062.

De Serrano, A. R., Hughes, K. A., & Rodd, F. H. (2021). Paternal exposure to a common pharmaceutical (Ritalin) has transgenerational effects on the behaviour of Trinidadian guppies. *Scientific Reports*, *11*, 1–13.

Deacy, W. W., Erlenbach, J. A., Leacock, W. B., Stanford, J. A., Robbins, C. T., & Armstrong, J. B. (2018). Phenological tracking associated with increased salmon consumption by brown bears. *Scientific Reports*, *8*, 1–9.

Ganzach, Y., (1998), Nonlinearity, multicollinearity and the probability of type II error in detecting interaction. *Journal of Management, 24*, 615–622.

Garamszegi, L. Z. (2011). Information-theoretic approaches to statistical analysis in behavioural ecology: An introduction. *Behavioral Ecology and Sociobiology*, *65*, 1–11.

Gauy, A. C. dos S., Boscolo, C. N. P., & Gonçalves-de-Freitas, E. (2018). Less water renewal reduces effects on social aggression of the cichlid *Pterophyllum scalare*. *Applied Animal Behaviour Science*, *198*, 121–126.

Glavaschi, A., Cattelan, S., Grapputo, A., & Pilastro, A. (2020). Imminent risk of predation reduces the relative strength of postcopulatory sexual selection in the guppy: Predation risk and sperm competition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*, 20200076.

Green, S. B. (1991). How many subjects does it take toto do a regression analysis? *Multivariate Behavioral Research*, *26*, 499–510.

Hansen, T. J., Fjelldal, P. G., Folkedal, O., Vågseth, T., & Oppedal, F. (2017). Effects of light source and intensity on sexual maturation, growth and swimming behaviour of Atlantic salmon in sea cages. *Aquaculture Environment Interactions*, *9*, 193–204.

Hansen, W. B., & Collins, L. M., (2004). Seven ways to increase power without increasing n. *Biologia Centrali-Americaa, 2,* 184-195.

Hertz, P.E. (1983). Homage to Santa Anita: Thermal sensitivity of sprint speed in agamid lizards. *Evolution, 37*, 1075–1084.

Hurlbert, S.H., & Stuart H. (2014). Pseudoreplication and the design of ecological field experiments*. Ecological Society of America, 54*, 187–211.

Ito, M. H., Yamaguchi, M., & Kutsukake, N. (2018). Redirected aggression as a conflict management tactic in the social cichlid fish julidochromis regani. *Proceedings of the Royal Society B: Biological Sciences*, *285*, 1 – 9.

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

Kasper, C., Colombo, M., Aubin-Horth, N., & Taborsky, B. (2018). Brain activation patterns following a cooperation opportunity in a highly social cichlid fish. *Physiology and Behavior*, *195*, 37–47.

Konietschke, F., Bösiger, S., Brunner, E., & Hothorn, L. A. (2013). Are multiple contrast tests superior to the ANOVA? *International Journal of Biostatistics*, *9*, 63–73.

Kreyling, J., Schweiger, A. H., Bahn, M., Ineson, P., Migliavacca, M., Morel-Journel, T., Christiansen, J. R., Schtickzelle, N., & Larsen, K. S. (2018). To replicate, or not to replicate – that is the question: how to tackle nonlinear responses in ecological experiments. *Ecology Letters*, *21*, 1629–1638.

Lakens, D., & Albers, C. J. (2017). When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias. *Journal of Experimental Social Psychology, 74*, 187–195.

Lim, C., Sen, P. K., & Peddada, S. D. (2013). Robust nonlinear regression in applications. *Journal of the Indian Society of Agricultural Statistics. Indian Society of Agricultural Statistics*, *67*, 215–234.

Lucon-Xiccato, T., & Bisazza, A. (2017). Sex differences in spatial abilities and cognitive flexibility in the guppy. *Animal Behaviour*, *123*, 53–60.

Magris, M., Cardozo, G., Santi, F., Devigili, A., & Pilastro, A. (2017). Artificial insemination unveils a first-male fertilization advantage in the guppy. *Animal Behaviour*, *131*, 45–55.

McCrum-Gardner, E. (2010). Sample size and power calculations made simple. *International Journal of Therapy and Rehabilitation*, *17*, 10–14.

McDonnell, L. H., Reemeyer, J. E., & Chapman, L. J. (2019). Independent and interactive effects of long-term exposure to hypoxia and elevated water temperature on behavior and thermal tolerance of an equatorial cichlid. *Physiological and Biochemical Zoology*, *92*, 253–265.

Miletto Petrazzini, M. E., Bisazza, A., Agrillo, C., & Lucon-Xiccato, T. (2017). Sex differences in discrimination reversal learning in the guppy. *Animal Cognition*, *20*, 1081–1091.

Nelder, A. J. A., Wedderburn, R. W. M., (1972). Generalized Linear Models. *Wiley for the Royal Statistical Society. 135*, 370-384.

O'Neill, S. J., Williamson, J. E., Tosetto, L., & Brown, C. (2018). Effects of acclimatisation on behavioural repeatability in two behaviour assays of the guppy *Poecilia reticulata*. *Behavioral Ecology and Sociobiology*, *72*:166.

Oldham, T., Dempster, T., Fosse, J. O., & Oppedal, F. (2017). Oxygen gradients affect behaviour of caged Atlantic salmon *Salmo salar*. *Aquaculture Environment Interactions*, *9*, 145–153.

Peres-Neto, P. R., & Olden, J. D. (2001). Assessing the robustness of randomization tests: Examples from behavioural studies. *Animal Behaviour*, *61*, 79–86.

Peterman, R.M.,& Randall M. (1990). Statistical power analysis can improve fisheries research and management. *Canadian Journal of Fishes and Aquatic Sciences, 47,* 2-15.

Petraitis P.S. (1998). How can we compare the importance of ecological processes if we never ask, “compared to what?” In: Resetarits Jr WJ and Bernardo J (Eds). Experimental ecology: issues and perspectives. New York, NY: Oxford University Press.

Plonsky, L., & Oswald, F. L. (2017). Multiple regression as a flexible alternative to ANOVA in L2 research. *Studies in Second Language Acquisition*, *39*, 579–592.

Richter, T. (2006). What is wrong with ANOVA and multiple regression? Analyzing sentence reading times with hierarchical linear models. *Discourse Processes*, *41*, 221–250.

Rosengren, M., Thörnqvist, P. O., Johnsson, J. I., Sandblom, E., Winberg, S., & Sundell, K. (2017). High risk no gain-metabolic performance of hatchery reared Atlantic salmon smolts, effects of nest emergence time, hypoxia avoidance behaviour and size. *Physiology and Behavior*, *175*, 104–112.

Solomon-Lane, T. K., & Hofmann, H. A. (2019). Early-life social environment alters juvenile behavior and neuroendocrine function in a highly social cichlid fish. *Hormones and Behavior*, *115*, 104552.

Sullivan, G. M., & Feinn, R. (2012). Using effect size—or why the P value is not enough. *Journal of Graduate Medical Education*, *4*, 279–282.

Sumpter, D. J. T., Szorkovszky, A., Kotrschal, A., Kolm, N., & Herbert-Read, J. E. (2018). Using activity and sociability to characterize collective motion. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*, 20170015.

Tomkins, P., Saaristo, M., Bertram, M. G., Tomkins, R. B., Allinson, M., & Wong, B. B. M. (2017). The agricultural contaminant 17B-trenbolone disrupts male-male competition in the guppy (*Poecilia reticulata*). *Chemosphere*, *187*, 286–293.

Walsh, S. L., Preston, K. L., Stitzer, M. L., Cone, E. J., & Bigelow, G. E. (1994). Clinical pharmacology of buprenorphine: ceiling effects at high doses. *Clinical Pharmacology & Therapeutics*, *55*, 569–580.

Wang, L., Zhang, Z., McArdle, J. J., & Salthouse, T. A. (2008). Investigating ceiling effects in longitudinal data analysis. *Multivariate Behavioral Research*, *43*, 476–496.

Wang, Y. A., Sparks, J., Gonzales, J. E., Hess, Y. D., & Ledgerwood, A. (2017). Using independent covariates in experimental designs: Quantifying the trade-off between power boost and Type I error inflation. *Journal of Experimental Social Psychology*, *72*, 118–124.

Wissuwa, M. (2003). How do plants achieve tolerance to phosphorus deficiency? small causes with big effects. *Plant Physiology*, *133*, 1947–1958.

Wolcott, H. L., Ojanguren, A. F., & Barbosa, M. (2017). The effects of familiarity on escape responses in the Trinidadian guppy (*Poecilia reticulata*). *PeerJ*, *2017*, 1–17.

Wood, S.N. Generalized Additive Models, An Introduction with R, 2nd edition. Chapman and Hall/CRC, New York.

Zar, H.H. (2022). Biostatistical Analysis, 5th edition. Pearson, London.

Ziegelbecker, A., Remele, K., Pfeifhofer, H. W., & Sefc, K. M. (2021). Wasteful carotenoid coloration and its effects on territorial behavior in a cichlid fish. *Hydrobiologia*, *848*, 3683-3698.

Zukoshi, R., Savelli, I., & Novales Flamarique, I. (2018). Foraging performance of two fishes, the threespine stickleback and the Cumaná guppy, under different light backgrounds. *Vision Research*, *145*, 31–38.

Ware, M. (2008). Peer review: benefits, perceptions and alternatives. *Publishing Research Consortium*, 20.

Scargle, J. D. (1999). *Publication Bias (The "File-Drawer Problem") in Scientific Inference*.

Dong, P., Loh, M., & Mondry, A. (2005). The "impact factor" revisited. *Biomedical Digital Libraries*, *2*, 1–8.

Preacher, K. J., MacCallum, R. C., Rucker, D. D., & Nicewander, W. A. (2005). Use of the extreme groups approach: A critical reexamination and new recommendations. *Psychological Methods*, *10*(2), 178–192.

Guisan, A., Edwards, T. C., & Hastie, T. (2002). Generalized linear and generalized additive models in studies of species distributions: Setting the scene. *Ecological Modelling*, *157*(2–3), 89–100.

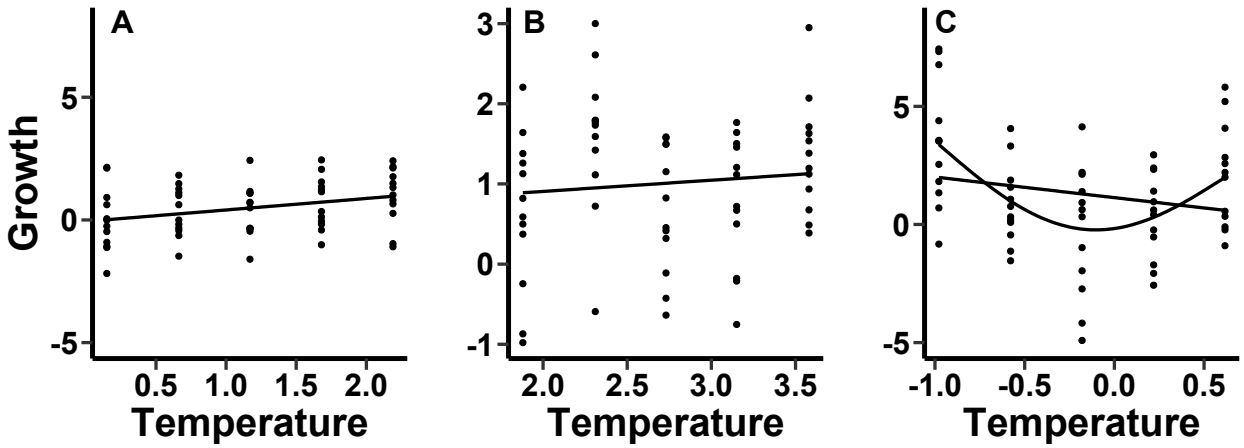

Figure 1 - Different result outcomes of a fictional experiment (simulated data) assessing the influence of temperature on fish growth. A) a case in which both ANOVA and regression are appropriate but the latter has greater statistical power (p-value = 0.238 and 0.020 for ANOVA and regression, respectively); B) a case in which both ANOVA and regression are appropriate but the former has greater statistical power (p-value = 0.025 and 0.478 for ANOVA and regression, respectively); C) a case in which only ANOVA is appropriate due to the non-linear nature of the relationship (p-value = 0.001 and 0.138 for ANOVA and regression, respectively). When fitting a non-linear regression model (see main text) to the relationship, the p-value = 0.0001.

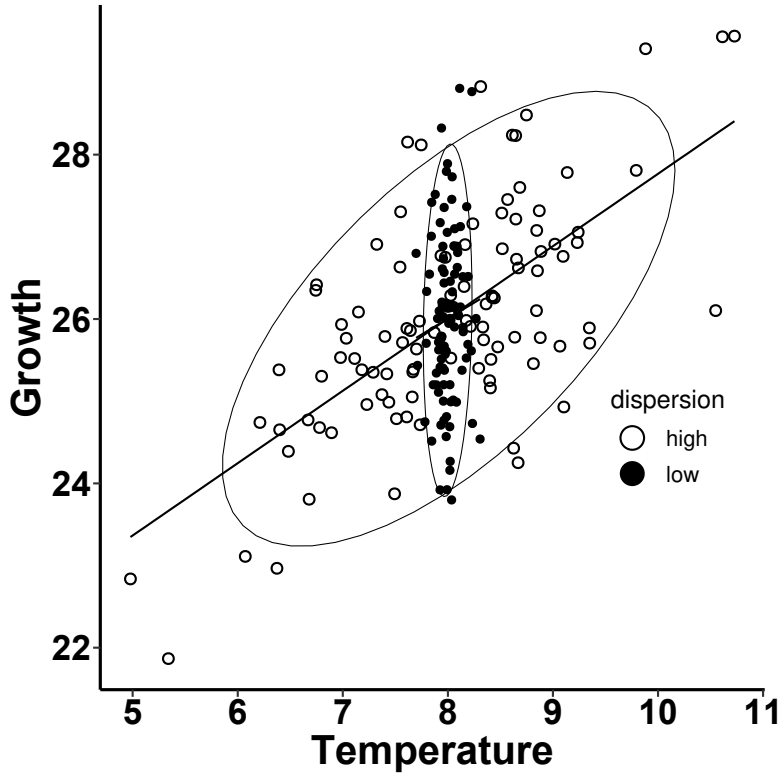

Figure 2. Figure showing regression models for two different dispersion (variance) values for a fictional field study (simulated data) assessing the influence of temperature on fish growth. The high dispersion and low dispersion series have about the same slopes (0.8804 and 0.7755, respectively), but their p-values (statistical power) differ dramatically (p-value = 2.602e-14 and 0.3948 for high and low dispersions, respectively). This illustrates that the dispersion of the predictor dramatically affects the standard error of the slopes, thus influencing p-value (standard error of the slope = 0.09864 and 0.9073 for high and low dispersions, respectively).

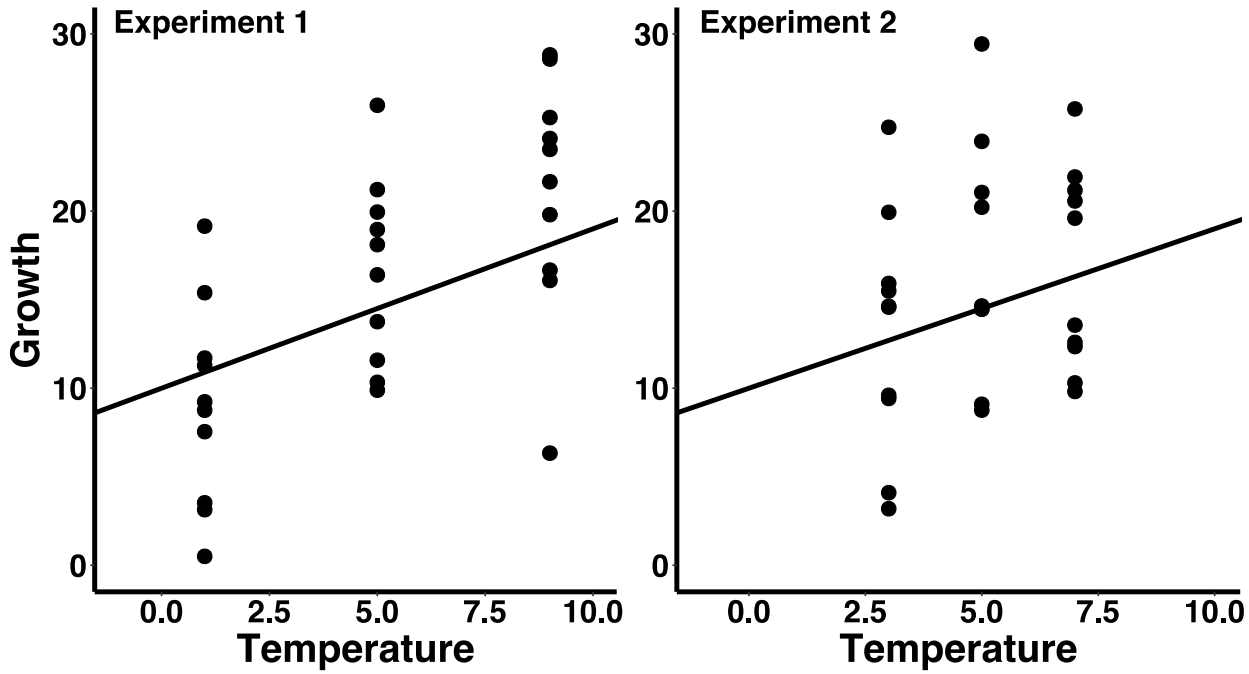

Figure 3. Illustration of how treatment dispersion affects both effect size and power in ANOVA. Each data (graph) represents an experimental outcome for a fictional experiment study (simulated data) assessing the influence of temperature on fish growth. Both data sets were generated to have the same regression intercepts and slopes, but their treatment dispersal varies dramatically (variance of treatments is 16 and 4 for the high dispersion and low dispersion treatment, respectively). Table 1 contains the ANOVA and effect sizes for each experiment. The effect of temperature on growth can only be detected in experiment 1.

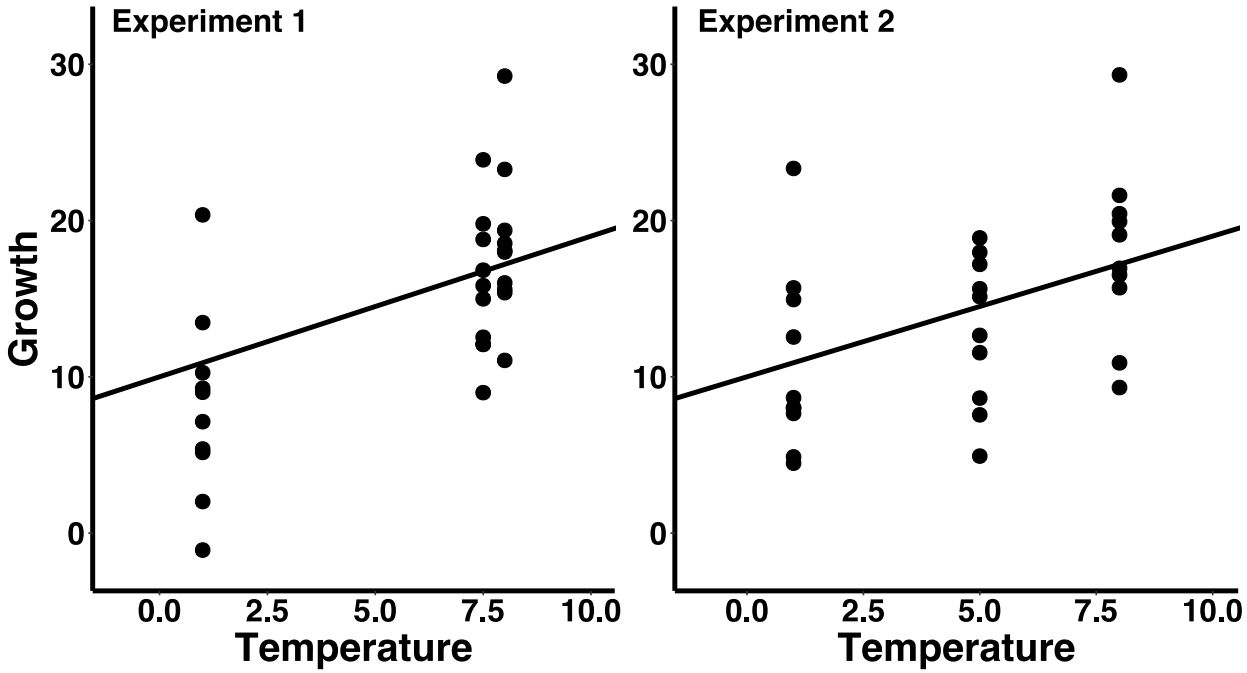

Figure 4. Illustration of how treatment dispersion affects both effect size and power in ANOVA. Each data (graph) represents an experimental outcome for a fictional experiment study (simulated data) assessing the influence of temperature on fish growth. Both data sets were generated to have the same regression intercepts and slopes, as well as the same temperature range, but their treatment dispersal varies (variance of treatments is 15.25 and 12.33 for the high dispersion and low dispersion treatment, respectively). Table 2 contains the ANOVA and effect sizes for each experiment. The effect of temperature on growth is much stronger in experiment 1.

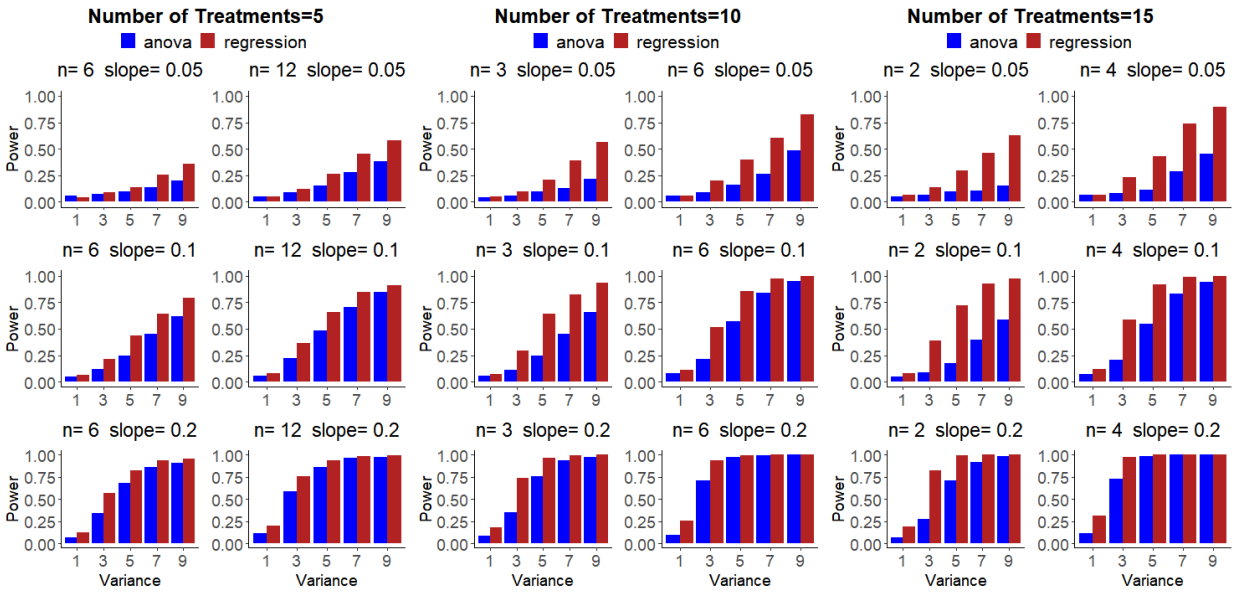

Figure 5. Power analysis (rejection rates over 1000 simulated data in each combination) comparing ANOVA versus regression for different data simulated using different regression slopes (either 0.05, 0.1 or 0.2; see methods) and treatment dispersions (variance) as a function of number of treatments (levels of groups) and the number of replicates per level (n). The total number of observations is either 30 or 60 observations and these are a function of the number of treatments × n (number of replicates per treatment) in any given plot.

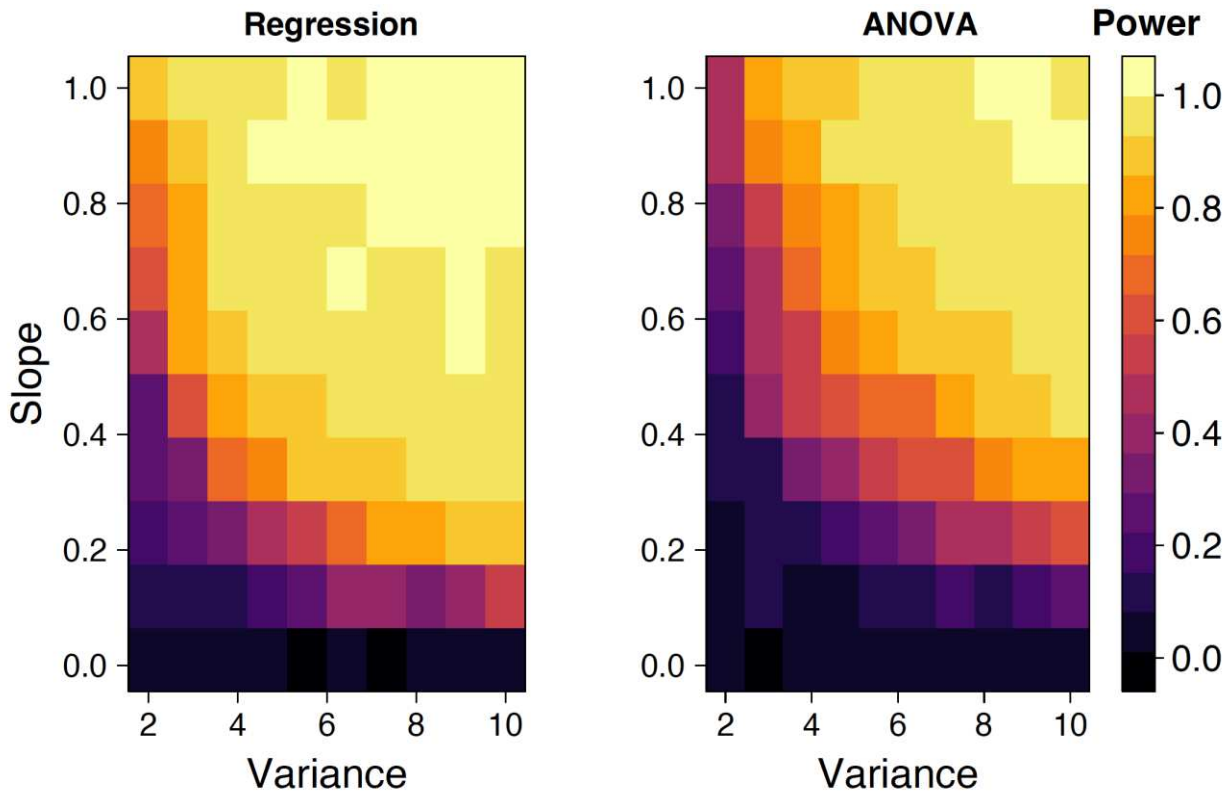

Figure 6. Power analysis (rejection rates over 1000 simulated data in each combination) comparing ANOVA versus regression for different data simulated using various regression slopes and treatment dispersions (variance). Number of treatment levels was fixed at 5 and number of replicates per treatment (n) was fixed at 6. The total number of observations is either 30 or 60 observations and these are a function of the number of treatments x n in any given plot.

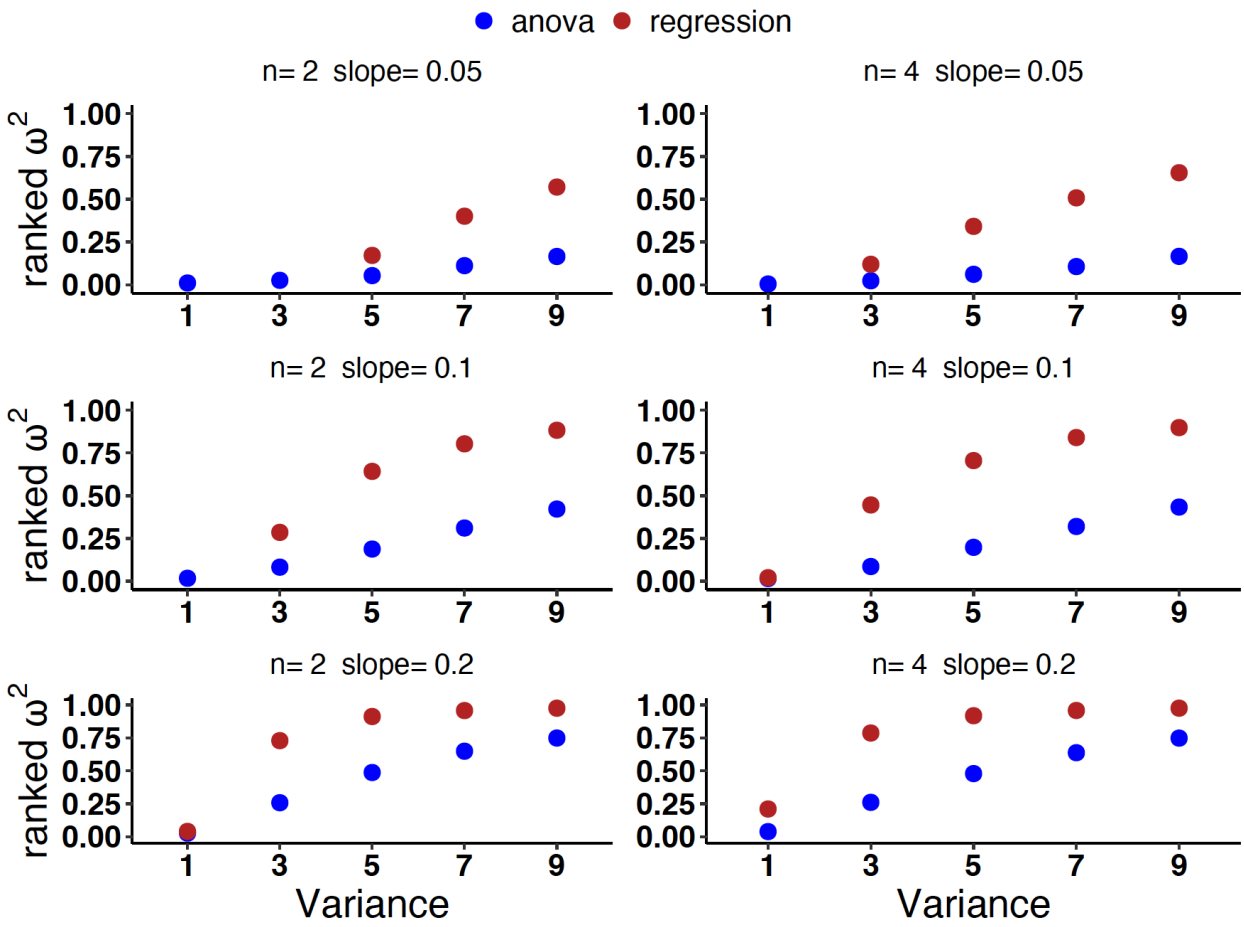

Figure 7. Average effect sizes (based on 1000 simulated data in each combination) comparing ANOVA versus regression for different data simulated using various regression slopes and treatment dispersions (variance). Here the number of treatment levels was set to 10 across all simulations and replicates per level (n) were either 3 or 6 observations.

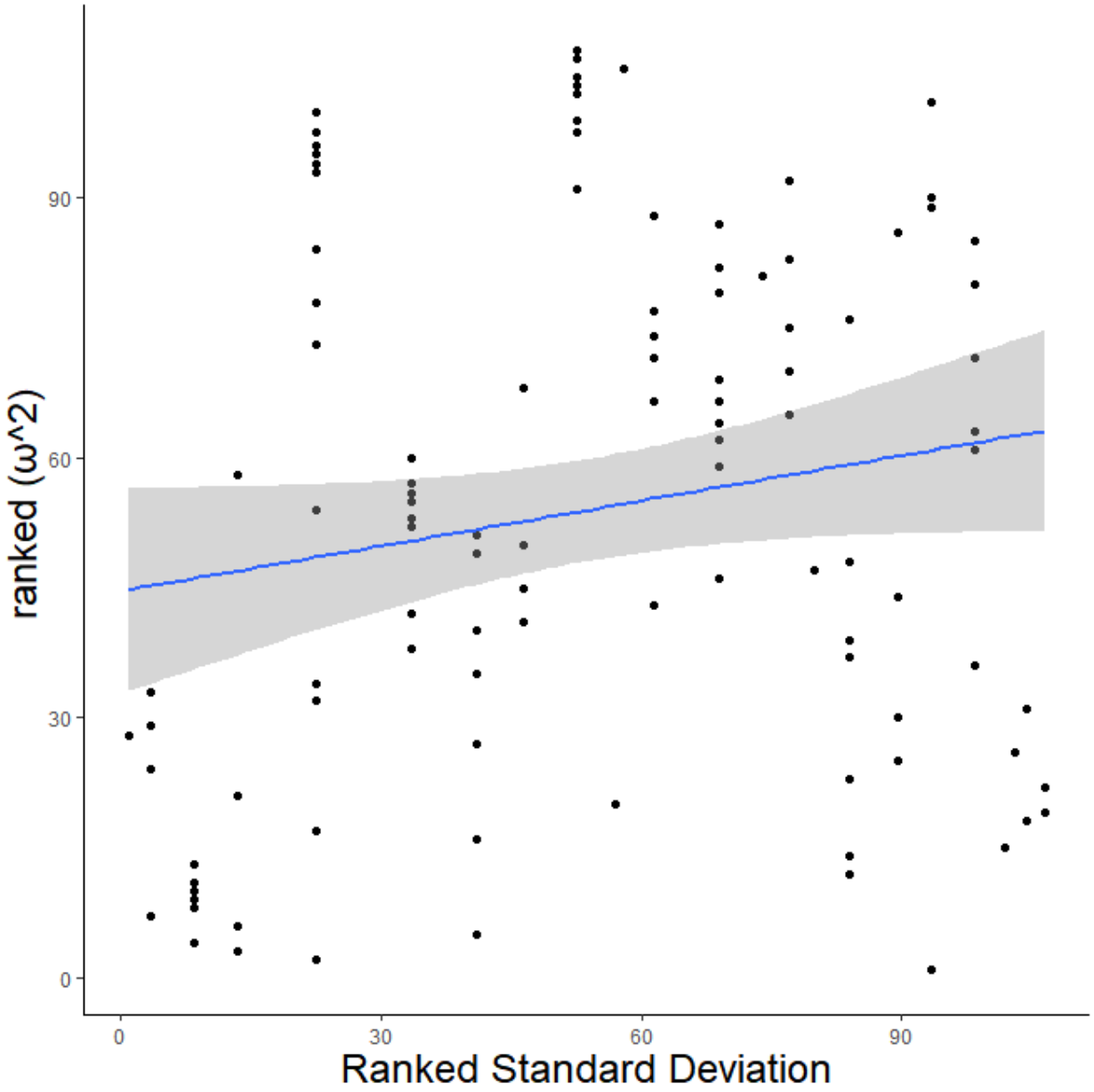

Figure 8. Meta-analytical results based on empirical experimental results using ANOVA on several studies on fish behaviour. Effect size was calculated as partial omega-squared (ω^2). Values were ranked to improve regression assumptions. Ranked standard deviation was used to measure treatment dispersion across treatments representing different variables and scales. While exhibiting a negative trend (slope of 0.55) these results are not significant (t = 0.76, p = 0.45, n = 107; see Table 3).

Table 1. Results of an ANOVA conducted on the fictional simulated experiments 1 and 2 in Figure 3.

	DF	SS	MSS	F	P	Cohen's *f*
Experiment 1						
Temperature	2	743.41	371.70	10.524	0.0004175	0.89
Residuals	27	953.58	35.32			
Experiment 2						
Temperature	2	94.5	47.249	4.5523	0.3211	0.30
Residuals	27	1076.4	39.866			

Table 2. Results of an ANOVA conducted on the fictional simulated experiments 1 and 2 in Figure 4.

	DF	SS	MSS	F	P	Cohen's f
Experiment 1						
Temperature	2	571.61	285.81	10.719	0.0003746	0.89
Residuals	27	719.95	26.67			
Experiment 2						
Temperature	2	268.90	134.45	4.5523	0.01978	0.58
Residuals	27	797.43	29.54			

Table 3. Results of Mixed Effects Linear Regression model of the effect of treatment dispersion on effect size. Pseudo-R^2 (fixed effects) = 0.01. Pseudo-R^2 (total) = 0.55.

	Slope	Standard Error	Degrees of Freedom	t-value	p-value	Standard deviation
Fixed Effect (Treatment dispersion)	0.10	0.13	78.70	0.76	0.45	
Intercept	43.28	9.25	34.09	4.19	< 0.0001	
Random Effect (Paper; intercept)						23.81
Residual						21.60