# More individuals or more groups? Incorporating sampling effort, statistical power, and model accuracy when designing experiments

**Lynch C.M.[1], Starkey M.[2], Montgomery, D.[2], Pavlic, T.P.[1,2], Mizumoto N.[3]**

[1]School of Life Sciences, Arizona State University, Tempe, AZ, USA 85287
[2]School of Computing and Augmented Intelligence, Arizona State University, Tempe, AZ, USA 85287
[3]Department of Entomology & Plant Pathology, Auburn University, Auburn, AL, USA, 36849

Corresponding author: Colin Lynch
cmlynch2@asu.edu
Arizona State University
School of Life Sciences
PO Box 874501
Tempe, AZ 85287-4501

**Abstract**

When testing differences between populations, researchers must balance sampling effort between repeated measures (e.g., sampling within the same population or field site) and independent replications (sampling across populations or field sites). Linear Mixed-effects Models (LMMs) are powerful tools for analyzing such data but depend on accurately estimating within- and across-group variance. Poorly chosen sampling strategies can lead to inflated type-I error rates and reduced statistical power. Using a social-insect example, we examined how sampling strategies impact LMM performance. For nested experiments—where different colonies receive different factor levels—sampling choices significantly affected type-I and type-II errors due to poor variance estimation. Conversely, crossed (full-factorial) designs, where each colony experienced all factor levels, were less sensitive to sampling effort allocation. Generally, increasing independent replications (e.g., sampling more colonies) improved nested experiment accuracy, though this often incurs higher costs than adding repeated measures. Our literature survey revealed social-insect studies sample 2.94 times more repeated measures than independent replications. To address this, we developed optimization protocols that integrate false-positive rates, sampling effort, and power analyses to design cost-effective experiments. These protocols provide practical guidance for balancing repeated measures and independent replications to achieve robust and economical experimental designs across multi-scale studies.

**Keywords**
Design of experiments, error rates, linear mixed model, power analysis, random effects, sampling techniques, social insects, repeated measures

## 1. Introduction

Datasets in ecology and evolutionary biology have multi-leveled structures which can drastically affect the outcomes of measurements made at each level of the experiment (Harrison, 2014). These datasets include covariates for various aspects of biological or experimental organization such as genotypes, species, temporal periods, and observer effects, the presence of which can alter the interpretation of a statistical test (Bolker et al., 2009; Schielzeth et al., 2020). Properly accounting for these structures in an experimental design is necessary and can enhance the validity of statistical inference (Quinn & Keough, 2002). For example, it is often necessary to sample across multiple scales of biological organization, as many populations are necessarily nested within others (Marshall, 2024). Furthermore, even when such nesting is not present, taking repeated measures of a single individual can, if properly accounted for (e.g., as random effects in linear mixed models), effectively reduce variance across individuals and thus provides an alternative to sampling more individuals to increase statistical power. There are therefore at least two degrees of sampling freedom in practical experimental design: *how many independent samples* to take across these levels versus *how many repeated measures* to take within these levels. Power analyses can help to determine how many independent replications of an experiment are necessary to detect a given effect size (Jennions & Møller, 2003), but power analyses do not provide any guidance about the connection between the number of repeated measures and statistical power. Furthermore, the parametric tests that have been developed for complex experimental designs such as those involving random effects often require subtle assumptions about the data that are not always carefully checked (Uttley, 2019). For example, even when all of the normality assumptions of a linear mixed model are met, good performance of the model requires it to make accurate estimates of within- and between-group variance, which itself adds a second implicit upward pressure on sample size to ensure that the designed significance level matches the actual type-I error (which is normally taken for granted so long as more conspicuous parametric assumptions are met, such as normality assumptions for linear models). Complicating matters further, the effort required to make another independent sample may be similar, much greater, or much less than the effort required for another repeated measure, and practical experimental designs must consider error rates but also implementation feasibility. In this study, we investigate optimal sampling strategies, asking how many samples should be taken within some biologically defined group (where measurements are potentially dependent on another, as in repeated measurements of the same individual) as well as how many groups should be sampled from (where groups are independent of one another, as in measurements made of two different unrelated individuals) given different experimental constraints on sampling budget.

Although the question of how to sample across different biological strata is ubiquitous in the biological sciences, we frame this study for a concrete example in the context of social-insect science. Here, insects are necessarily members of a larger colony, which is a social group usually composed of a few reproductives and their non-reproductive offspring (Hölldobler & Wilson, 2009). Colony members therefore will often share the same genetic and

environmental background, and two individuals from the same colony cannot be viewed as two independent samples when making inferences scoped to a larger population of colonies. The effect of the colony is non-negligible, as colonies can largely vary in behavior and shape colony personality (Jandt et al., 2014; Wright et al., 2019) due to differences in their genetic makeup and developmental state (Bengston & Jandt, 2014). That said, generally speaking, individuals within a colony are not genetic clones nor have they shared exactly the same developmental history, and so there is significant variance across nestmates as well. Consequently, testing the effect of an experimental manipulation on a population of colonies is benefitted not only by sampling a diversity of colonies but also by sampling several individuals within each colony and controlling for the effect of the colony. The discrete nature of these colonies and the inherent differences between colonies make them a useful test case for studying the effects of sampling across populations generally. We emphasize that although our study was parameterized for sampling across colonies, the general method could be used for sampling across other levels such as repeated measurements of individuals. Indeed, we encode the colony as a random effect in a linear mixed model (LMM), which can be done for many of the structures present in ecological datasets (Bolker et al., 2009; Udino et al., 2017; Carere et al., 2018).

We introduce three broad strategies for sampling across these colonies that we refer to as: (a) depth collection, sampling a large number of individuals from a small number of colonies; (b) breadth collection, sampling a small number of individuals from a large number of colonies; or (c) balanced collection, sampling as many individuals per colony as the number of colonies. Collecting a new colony from a field or rearing a newly established laboratory colony is more costly than measuring another single individual from each of the already established nests. Thus, for a given sample budget, depth collection is likely more practical than breadth or balanced collection. However, an insufficient number of colonies may also introduce some unintentional bias to statistical analyses. Thus, as in the choice to invest more in additional samples or better microscopy equipment (Gundersen & Østerby, 1981), budget induces a non-trivial tradeoff among different methods for reducing variance (and increasing statistical power) in a study. As these three strategies are not equivalent in terms of sampling costs or performance of statistical tests, the advantages and disadvantages of each method need formal quantification within social-insect science like what has been done in other biological fields (Baker et al., 2021; Frank et al., 2023).

The performance of different sampling strategies can be measured by the ability to minimize false negative rates (type-II error) and false positive rates (type-I error) in statistical tests. Power analyses calculate the former, as power is the complement of the false negative rate. Power analyses for LMMs use simulation-based approaches to estimate power (Johnson et al., 2014; Kain et al., 2015; Green & MacLeod 2016). These approaches can be performed over multiple hierarchical levels of observation. For instance, work in psychology leverages these methods to calculate power contours across independent subjects and repeated trials per subject (Baker et al., 2021; Chen et al., 2021). However, these studies make the implicit assumption that the collection strategy does not inflate the true false-positive rate beyond the desired significance level, ignoring the possibility that the sampling approach may have parasitic effects that increase false-positive rates (e.g., Wang et al., 2017).

Here, we investigate the effectiveness of different sampling strategies to estimate the fixed effect of a two-leveled categorical variable (i.e., experimental treatment) on normally

distributed, synthetic data structured by colony. We take into account the asymmetry of sampling costs across levels. That is, we incorporate the fact that the effort to sample more colonies is not equal to that of the effort required to obtain more within-colony replicates. First, we survey the published literature on social insects to investigate their sampling strategies, finding the number of within-colony replicates and among-colony replicates for each study. Then, we compare the performance of depth, breadth, and balanced collection across different sample sizes with various free-parameter combinations. We also accounted for colony effects to be either nested or crossed depending on the experimental design (Schielzeth & Nakagawa, 2013). Crossed designs (or, equivalently, full factorial designs; Montgomery, 2020), ensure that each colony is assessed across levels of the fixed effect, and nested designs use different colonies for each level of the fixed effect. Finally, we performed a social-insect case study to demonstrate how the relative challenges of sampling within- and across-colonies could be integrated into an optimization framework focused on simultaneously minimizing sampling effort and error rates. This study provides an overview of distributing a budget of sampling effort across and within biologically relevant groups.

**2. Literature survey on sample size used in social-insect research**

To get an overview the latest trends in the sampling strategies and experimental designs in social insect research, we surveyed articles from volumes 65–68 (the latest issues when the survey was performed) of *Insectes Sociaux*, the journal of the International Union for the Study of Social Insects, which published articles exclusively on social insect research. Among 76 articles on these issues, we focused on 50 articles that empirically investigated the effect of a categorical variable (treatment) and explicitly reported their sample sizes. We extracted the number of within-colony replicates per colony per treatment (hereafter, $i$) and the number of colonies sampled per treatment (hereafter, $c$) from each paper with their experimental designs (nested or crossed). When $i$ and $c$ were not explicitly given, we calculated them by dividing the total number of colonies or replicates by the number of treatments, respectively. When there were multiple experimental variables within a single experiment, we counted the number of treatments as the total number of combinations of factor levels. Sample sizes were often unequal across treatment groups; in these cases, we used the arithmetic-mean sample size (Sakamoto et al., 2020). If there were multiple experiments in an article, we recorded only the first experiment listed in the methods section.

In social-insect research, it is often easier to obtain more within-colony replicates than to obtain more colonies because colonies consist of 100~1,000,000 individuals. Conversely, designs could be chosen to minimize different types of variability. For instance, in cases where the genetic makeup of queens is known to play a huge effect on the outcome of some response variable (Fjerdingstad et al., 2003), then one may choose to sample many colonies in order to account for this effect. We therefore defined $W \triangleq i/c$ to be the relative individual-to-colony sampling ratio in each article where $i$ is the number of within-colony replicates and $c$ the number of colonies. If $W > 1$, the study's authors chose to have more between-colony replicates than within-colony replicates (and vice versa for the case where $0 < W < 1$) either to minimize the cost of the experiment or to minimize variability at the colony level.

Regardless of the interpretation of $W$, we found that most papers used depth collection rather than breadth collection (35 out of 50 used depth collection, binomial test $p < 0.01$).

Overall, studies used more within-colony replicates than the number of colonies (Wilcoxon test statistic = 1,782, $p < 0.001$, median within colony samples = 13.66, median across colony samples = 9). Papers that utilized depth collection sampled 22 replicates from 4 colonies, while studies that used breadth collection sampled 4 individuals from 14 colonies. The relative effort of sampling more colonies to sampling more within-colony replicates was greater than 1 (median $W = 2.94$). The relative efforts were variable across taxonomic lineages of social insects (ants, bees, termites, wasps; Kruskal-Wallis test $\chi^2 = 11.38$, df = 3, $p < 0.01$), and all of them were larger than 1 (Ant $W = 2.92$; Bee $W = 5$; Wasp $W = 1.14$; Termite $W = 21.41$).

Crossed designs were used as often as nested designs (22 out of 50 used crossed designs, binomial test $p = 0.4799$). There was also no association between sampling strategies and experimental design (Fisher's exact test $p = 0.3673$). Generally, there was no difference in total sample size between nested and crossed designs (Wilcoxon test statistic = 352, $p = 0.3141$).

## 3. Evaluation of sampling strategies

### 3.1 Data generation

To compare the performance of various sampling strategies, we expanded the methodology of Bolker (2008). We simulated many experiments where the statistical goal was to compare the means of a control group and a treatment group. We ran simulations both with and without the effects of an experimental treatment (effect size = Δ). We assumed that the data followed a normal distribution (= sampling error: $\epsilon_j \sim N(0, \sigma^2_w)$ ) with additional errors following a normal distribution attributed to the original colony (= colony effect: $\gamma_j \sim N(0, \sigma^2_a)$ ). Thus, within- and among-colony variations are represented as $\sigma^2_w$ and $\sigma^2_a$, respectively. The *j*th observation of the response variable *y* was generated with the following function:

$$y_j = \Delta x_j + \gamma_j + \epsilon_j$$

where $x_j$ represents a dummy variable taking the value of 0 (control) or 1 (treatment). In the nested design, we drew different colony-specific values for each treatment to simulate different colonies between two treatments. In the crossed design, we drew the colony-specific values only once so that they were shared between treatments.

In each simulation, there were *n* samples, with *i* replicates from *c* colonies for each treatment (thus $n = i \times c$; where *i, c*, and *n* are all positive integers and the total sample size from the two treatments is $N = 2n$). Thus, sampling strategies are considered depth collection when $i > c$, balanced collection when $i = c$, and breadth collection when $i < c$. We explored the entire sampling space of *i* and *c* in the range between 1 and 20, except ($i$, $c$) = (1, 1), for a total of $20^2 - 1 = 399$ combinations. Our upper limit of 20 covered more than half of the sample sizes in the literature review (67%). For each combination of *i* and *c,* we investigated the effect of three parameters Δ, $\sigma_w$, and $\sigma_a$, ranging from 0 to 3 (Table 1), with $10^3 = 1{,}000$ combinations. Note that our statistical tests are not sensitive enough to detect small $\Delta < 0.3$ (Section S 1.1), and small effect sizes are generally negligible (Cohen, 2016). Instead of exploring values of $\Delta < 0.3$, we triplicated $\Delta = 0$ to emphasize the importance of avoiding false positives (Table 1). We repeated this process 50 times to count the number of true and false positives as well as the

number of true and false negatives. Then we repeated this process for both crossed designs and nested designs. Thus, we performed 399 × 1,000 × 50 × 2 = 39,900,000 simulations in total. Simulations were performed in MATLAB (version 2023a).

**Table 1:** Free parameter values for data generation. We tested all combinations of these parameters to evaluate different sampling strategies.

| Parameter | Range |
|---|---|
| Mean Shift (Δ) | ∈ {0, 0, 0, 0.3, 0.75, 1.2, 1.65, 2.1, 2.55, 3} |
| Within Colony Variance ($\sigma^2_w$) | ∈ {0, $(1/3)^2$, $(2/3)^2$, $1^2$, $(4/3)^2$, $(5/3)^2$, $2^2$, $(7/3)^2$, $(8/3)^2$, $3^2$} |
| Among Colony Variance ($\sigma^2_a$) | ∈ {0, $(1/3)^2$, $(2/3)^2$, $1^2$, $(4/3)^2$, $(5/3)^2$, $2^2$, $(7/3)^2$, $(8/3)^2$, $3^2$} |

### 3.2 Evaluation of sampling strategies

For each generated dataset, we fit a linear mixed model (LMM) that includes experimental treatment as a fixed effect and colony as a random effect (random intercept). In principle, these data could also be analyzed with a repeated-measures ANOVA. We chose instead to use an LMM as it is a more generic model for capturing sources of variance, and we expect that results that hold for the LMM will be qualitatively similar to results for an ANOVA. We tested for statistical significance of the fixed effect by calculating *p* values from the associated *F* statistic. We determined whether the model reached the correct conclusion for each simulated dataset. When $\Delta > 0.3$, we recorded $p < \alpha$ (where $\alpha = 0.05$) as a true positive and $p > \alpha$ as a false negative (type-II errors), Likewise, when $\Delta = 0$, we recorded $p < \alpha$ as a false positive (type-I errors) and $p > \alpha$ as a true negative result. We calculated the probability of obtaining a true positive (power) and true negative for each sampling strategy. We also computed the balanced accuracy as the average of true-positive and true-negative probability (Brodersen et al., 2010) to equally prioritize the minimization of both errors. Furthermore, we assessed how accurately the model estimated the parameters, Δ, $\sigma_w$, and $\sigma_a$ by calculating squared error (the squared difference between true and sample estimates of each parameter). Finally, we measured the excess kurtosis of the sample distributions to understand why some collection strategies had higher Δ errors (the squared difference between the true and estimated Δ values) than others. Negative values of excess kurtosis indicate that tails are heavier (outliers are more likely to occur) than those of normal distributions, so sample distributions with more outliers could pull the treatment means further apart, increasing estimated Δ.

We found that balanced accuracy and power did not depend on the sampling strategy for crossed designs. Rather, it depended purely on sample size (Fig. 1AC). Additionally, the probability of a true negative (probability of avoiding false positive) is constant regardless of sample size or sampling strategy (Fig. 1E). Crossed designs also achieved higher power, balanced accuracy, and true-negative rates than nested designs (Fig.1).

On the other hand, the performance of sampling strategies in nested designs depended on the sampling strategy (Fig. 1B, 1D, and 1F). Breadth collection generally achieved higher balanced accuracy than depth collection, that is, the overall performance increased with more among-colony replicates rather than more within-colony replicates (Fig. 1B). The exception was

at very small sample sizes, where the depth collection (with $c$ = 1) overperformed the breadth collection (with $i$ = 1) in higher balanced accuracy (Fig. 1A) and power (Fig. 1D). In this region, depth collection distinguished two treatments regardless of whether there were true differences, leading to higher power as well as lower true-negative rates (Fig. 1D and 1F). As balanced accuracy includes the true positive, this metric could be high when $c$ = 1 but not as high as power, which is not balanced by the true-negative rate. With larger sample sizes, breadth collection achieved higher balanced accuracy, power, and true-negative rates than depth collection (Fig. 1B, 1D, and 1F).

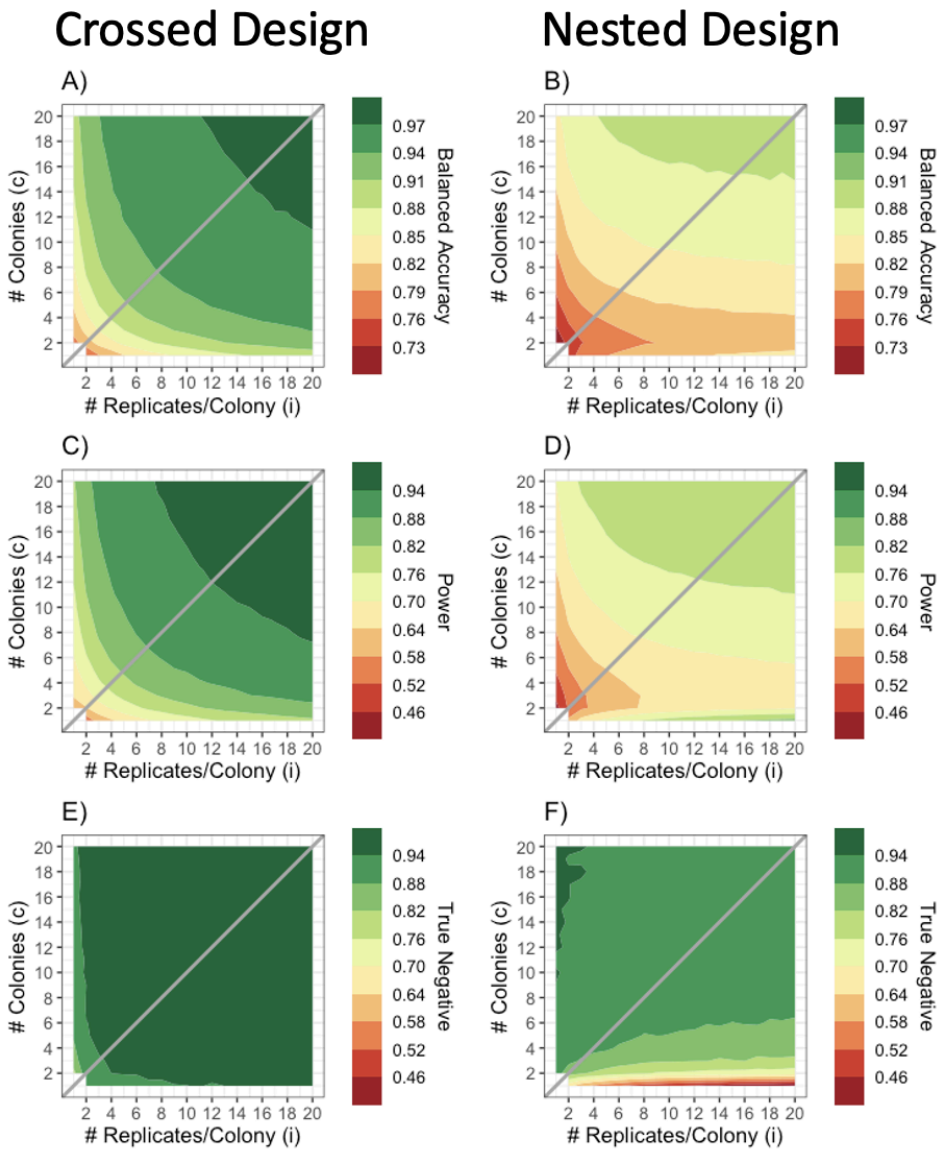


**Figure 1.** The relative advantage of depth and breadth collection across different sample sizes pooled across all parameter combinations. The performance was evaluated by either balanced accuracy (A, B), power (C, D), and the true negative rate (E, F). Crossed designs are in the first

column (A, C, E), and nested designs are in the second column (B, D, F). The solid diagonal lines represent balanced collection strategies ($i = c$). Above the line shows breadth collection strategies ($c > i$), and depth collection strategies lie below it ($c < i$).

The fact that the true negative rate is lower than 0.95 ( = 1 - $\alpha$ = 1 - 0.05) for nested designs even for large sample sizes is an indication that there is a violation in in the assumptions of our statistical models. This violation likely occurs from the poor variance estimates of the random effect, which in turn leads to a conflation of the treatment with the differences between colonies which is not properly controlled for in nested designs. In these designs, breadth collection was the more advantageous sampling method as it could estimate the total variances of their datasets more accurately. The total variance of a dataset was composed of within- and among-colony variance. At small sample sizes, breadth collection estimated among-colony variance better than depth collection (Fig. 2A), while depth collection estimated within-colony variance better than breadth collection (Fig. 2B). With increasing sample size, all sampling strategies could estimate both within and among colony variance more accurately, and thus converged into a similar level of estimation (Fig. 2AB). However, this pattern was distinct between within- and among-colony variance. Breadth and depth collection converged faster in estimating within-colony variance compared to estimating among-colony variance, but breadth collection converged faster to the true value of within-colony variance (Fig. 2B) than depth collection converged on across-colony variance (Fig. 2A).

Breadth collection also minimized effect-size error irrespective of sample sizes (Fig. 2C). This happened because the tails of the sample distributions for depth collected were heavier than those of breadth collection (it had a more negative value for excess kurtosis; Fig. 2D), which would draw out the sample means further apart from one another. When there was a true difference between treatments, the heavier tails (meaning that outliers are more likely to occur) would magnify the effect, making a true detection more likely (but the effect size estimate is less accurate). However, this would also create spurious differences between treatments even when there was no true difference, increasing the false positive rate (or decreasing the true negative rate; Fig. 2F). In other words, as the variance due to colonies is poorly estimated in nested designs, these differences are conflated with the treatment, so it sometimes appears that there is an effect of treatment when really there are just differences in the colonies tested between the two treatments.

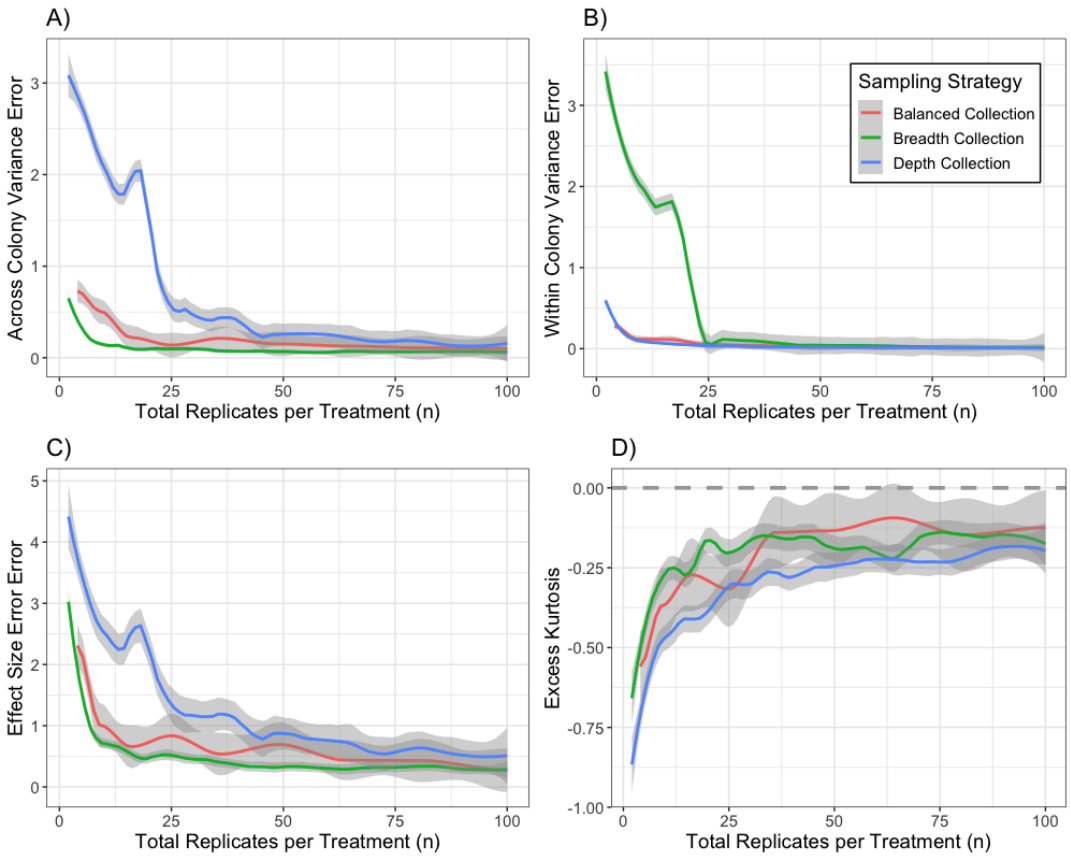


**Figure 2.** The performance of sampling strategies across different numbers of replicates per treatment (*n*) for nested designs. The solid lines show the LOESS regression for each collection type across *n*, and the shaded regions indicate 95% confidence intervals. In D, the horizontal dashed line indicates the excess kurtosis for a normal distribution. The x-axis is limited to *n* = 100, and data points were removed for visibility.

### 3.3 Optimizing sampling strategy with variable sampling efforts

Finally, we evaluated the sampling strategies after accounting for the relative difficulties of sampling at each level. We incorporated the asymmetry of replicates with the equation for sample size ($n = ic$) as follows:

$$E = ic + i(1/W - 1) + c(W - 1)$$

where $W$ sets the weight to sample among-colony replicates over sample within-colony replicates (Fig. 3). When $W = 1$, the sampling effort is constant across combinations of $i$ and $c$, given the same $n$, forming symmetric isoeffort contours (Fig. 3B, D). When $W > 1$, the among-colony sampling is more costly than within-colony sampling (and vice versa for when $0 < W < 1$). With $W \neq 1$, we obtained asymmetric isoeffort contours (Fig. 3C, D, E).

A)

| Strategy | # Replicates per Colony (i) | # Colonies (c) | Label |
|---|---|---|---|
| Breadth Collection | 1 | 16 | a |
| | 2 | 8 | b |
| Balanced Collection | 4 | 4 | c |
| Depth Collection | 8 | 2 | d |
| | 16 | 1 | e |

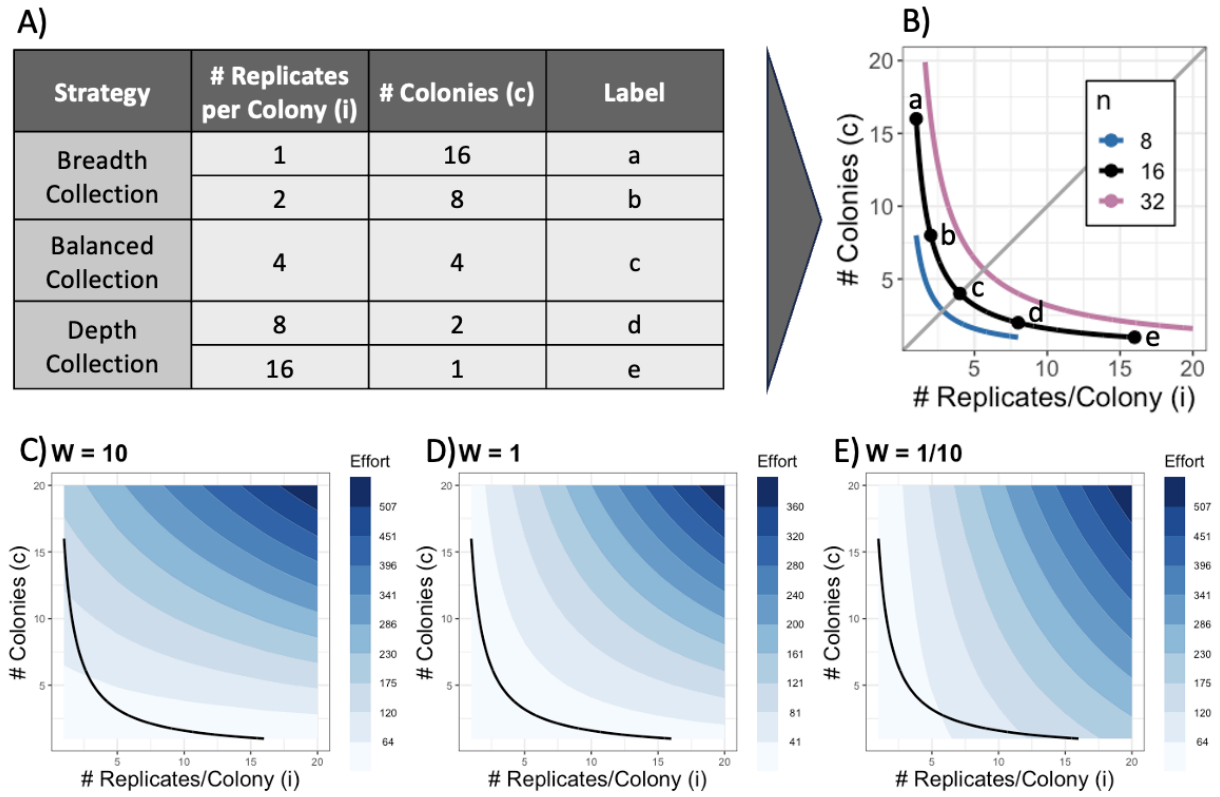


**Figure 3.** Required effort for different sampling strategies. (AB) Mapping sampling strategies on the isoeffort contours with $W$ = 1. When the treatment sample size has a fixed effort of $n$ = 16, there are 5 possible combinations of sampling strategies (a-e). This contour moves to the upper right when $n$ increases (green line) or to the bottom left when it decreases (pink line). (C, D, E) Isoeffort contours with differing relative efforts of sampling colonies vs sampling within colonies. When $W$ = 10, colonies are roughly 10 times harder to sample from than individuals within a colony, and so effort increases more when one moves along the y-axis of the plot rather than the x-axis. The opposite is true when $W$ = 1/10. Black lines on each panel represent the balanced case where $W$ = 1 and $n$ = 16, showing deviations in effort for $W$ > 1 and $W$ < 1.

In social insect research, depth collection is used more than breadth collection, where the number of within-colony replicates is 2.94 times more than the number of colonies in the median (section 2). It is usually impossible to infer how the authors selected the number of within- and among-colony replicates, and thus one should exercise caution when interpreting this value. Nevertheless, one possible implication is sampling more colonies is more costly than sampling more within-colony replicates. Here we used $W$ = 2.94 as a representative value of the sampling weight in the standard social insect research and investigated the consequences of choosing variable sampling techniques (all combinations of $i$ = 1, 2, …, 20 and $c$ = 1, 2, …, 20) as a case study.

We evaluated sampling strategies (combinations of $i$ and $c$) with how they can maximize performance given a fixed sampling effort ($E$) or how they can minimize sampling efforts given a certain acceptable performance. The performance was measured with statistical power and balanced accuracy. Balanced accuracy is the mean of the true positive and negative rates

(Alpaydin, 2020). Power thresholds (*P*) can therefore be converted to balanced accuracy thresholds (*B*) for two sided tests with:

$$B = \frac{P + (1 - \alpha)}{2}$$

where $\alpha$ is the significance level. We set the maximum acceptable effort (*E*) to 100, the minimum acceptable power to 0.8 (Crawley, 2007), and the minimum acceptable balanced accuracy was (0.8 + (1-0.05))/2 = 0.875. We used the parameters $\Delta = 0$ or 2.55 and $\sigma_w = \sigma_a = 2$. All statistical tests were performed in R (R Core Team, 2021) using the stats, dunn.test, tidyverse, FSA, and TTR packages.

In crossed designs, the performance of sampling is consistent across combinations of *i* and *c* with a given *n* (Fig. 1A, 1C, and 1E). When the upper limit of E was 100 (Fig. 4A, 4B, gold circle), (*i, c*) = (16, 1) maximized power. Alternatively, (*i, c*) = (7, 1) minimized E to achieve a power of at least 0.8 (Fig. 4A, 4B, gold triangle). The results are similar for balanced accuracy, where (16, 1) maximized balanced accuracy when E had an upper limit at 100 (Fig. 4C, 4D, gold circle). (5, 1) minimized E to achieve a balanced accuracy of 0.9 (Fig. 4C, 4D, gold triangle). Depth collection was selected as it has lower effort for the same levels of power and balanced accuracy.

In nested designs, the breadth collection was sometimes preferred over the depth collection despite $W > 1$. When we focused on power, breadth collection, (*i, c*) = (4, 20), maximized power when the effort was limited to 100 (Fig. 4E, 4F, gold circle), while depth collection with (4, 1), minimized E to achieve a power of 0.8 (Fig. 4E, 4F, gold triangle). However, this is problematic because the strategy (4, 1) fell into the region where the false-positive rate was higher than the conventional threshold of 0.05 (Fig. 1F). This did not occur when we used balanced accuracy as the performance metric. The optimal strategy was (3, 5) rather than (4, 1), showing that breadth collection can help avoid regions with high false-positive rates (Fig. 4G, 4H, gold triangle). The optimal strategy for maximizing balanced accuracy given the E limit was similar to that of the power estimate at (4, 15) (Fig. 4G, 4H, gold circle).

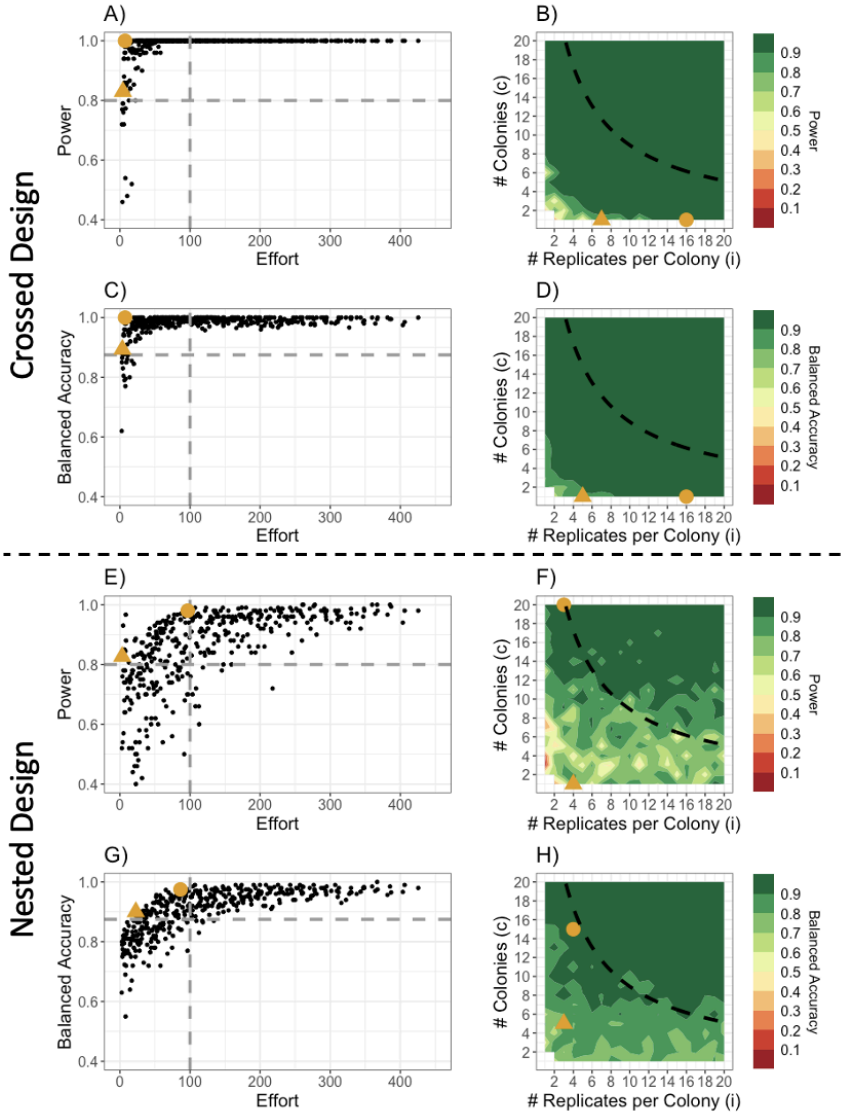


**Figure 4.** Optimal sampling strategies after accounting for effort. Plots in the left column show the relationships between effort and power (A, C) or balanced accuracy (B, D) for nested designs (A, B) and crossed designs (C, D) with parameters W = 2.94, Δ = 0 or 2.55 and $\sigma_w$, $\sigma_a$ = 2. The vertical gray dashed lines indicate effort = 100 while the horizontal dashed lines show the minimal acceptable power (= 0.8) or balanced accuracy (= 0.9). The circle is the strategy that maximizes power given an upper limit on effort. The triangle is the strategy that minimizes effort given a lower limit on power or balanced accuracy. These optimal strategies correspond to the circle and triangle in the right column, which show the sampling space for each design. The contours in the heat-maps correspond to the same level of effort as the vertical lines on the scatterplots.

## 4. Discussion

Our results highlight the gap between ideal sampling strategies and the real experimental practice used in social insect research. We found that previous studies

predominantly used depth collection rather than breadth collection, regardless of whether the experiment used a nested or crossed design. There was also no difference in sample sizes between studies that used nested or crossed designs, despite nested designs requiring larger sample sizes to account for additional variation (Lazic, 2018). However, we found that breadth collection can more effectively test experimental hypotheses in nested designs (Fig. 1). The advantage of breadth collection in nested design experiments is maintained even after accounting for the likely scenario where sampling colonies is more difficult or less desirable than sampling within colonies (Fig. 4). Thus, social insect researchers can profit from sampling additional colonies when they need to use nested experimental designs.

Breadth collection is preferred in nested designs because depth collection can lead to a high false-positive rate at small sample sizes. For small sample sizes, balanced accuracy was higher for depth collection than breadth collection, but this is, in part, an artifact of the way balanced accuracy is calculated. Balanced accuracy incorporates both true positive and true negative rates, and since scientists accept false negative rates (typically $\beta = 0.2$) four times higher than false positive rates (typically $\alpha = 0.05$), balanced accuracy is not sensitive to small deviations in the false-positive rate. These false positives arise because sample distributions have wider tails, resembling t-distributions. Thus, the difference in sample means between control and treatment groups can be artificially inflated, resulting in not only magnifying a true signal (true-positive rate increases) but also creating a signal that does not actually exist (a false positive), especially when only one colony was sampled for each treatment. However, the increase of true positives for breadth collection often outweighs the true positives from depth collection for larger sample sizes. Therefore, the balanced accuracy of breadth collection is higher than depth collection when the sample size per treatment ($n$) is greater than 14 (Fig. 2B). Interestingly, the total sample size ($N$) at $n = 14$ is $2n = 28$, which is close to the threshold for transitioning from a t-test to a z-test, which is colloquially at $N = 30$ (Cochran & Cox, 1964). The value $n = 15$, then, could represent a lower limit on sample sizes for nested designs, as this avoids a region of the sampling space where the type-I error rate is highest.

When power is used as the performance metric, sampling only one colony could be regarded as an optimal strategy (Fig. 4A). This is because power analyses rely solely on false negative rates (Cohen, 2013), where false positive rate is assumed to be consistent as 0.05 (Devane et al., 2004; Royston & Sauerbrei, 2013; Aberson, 2019). Thus, the power analysis is problematic when the type-I error rates can be variable, e.g., with small sample sizes (Luke, 2017), multiple comparisons (Colquhoun, 2014), or nested data structures (Aarts et al., 2015). Our study emphasized that this problem is ubiquitous in nested designs, as they generate poor estimates of variance which results in conflating differences between groups with differences between treatments. Conversely, this problem does not occur for crossed designs, so power is a sufficient metric for these kinds of experiments. Regardless, it is critical to consider both type-I and type-II error rates to determine the sample size when the sampling effort is limited and highly asymmetric between within- and across-colony replicates. Furthermore, exploring different significance levels other than traditional 0.05 could be useful as this could ascribe different weights to the type-I and type-II error rates (one could also use the area under the receiver operating characteristic curve in machine learning to explore all significance levels; Bradley, 1997).

In this study, we found that most social insect studies sample more individuals per colony than colonies ($W > 1$). However, the exact value of $W$ should be considered on an individual basis, depending on the experimental context. For instance, honeybee researchers need to buy additional equipment for new hives, increasing the monetary cost of the study. A rare species is difficult to collect in the wild. On the other hand, $W$ could be smaller than one, e.g., when studying plenty of incipient colonies with inevitably small colony size, or when individual replicates from each colony is a subgroup (e.g., 1,000 individuals). Thus, optimization can be achieved by maximizing performance given a maximum amount of effort or minimizing effort given a required performance. Overall, these optimization frameworks show that breadth collection outperforms depth collection for nested designs, while depth collection outperforms breadth collection in crossed designs. We summarize these results in a flow chart (Fig. 5). Ultimately, we find that there is significant variance across colonies and if you're limited to giving each colony a single factor level treatment, then it is more important to add more colonies if you can. If a colony can experience all treatments, then the experimentalist can do what is easiest: sample more often within or across colonies. We emphasize that the 'colony' moniker can be replaced with any ecological group that is present within a study.

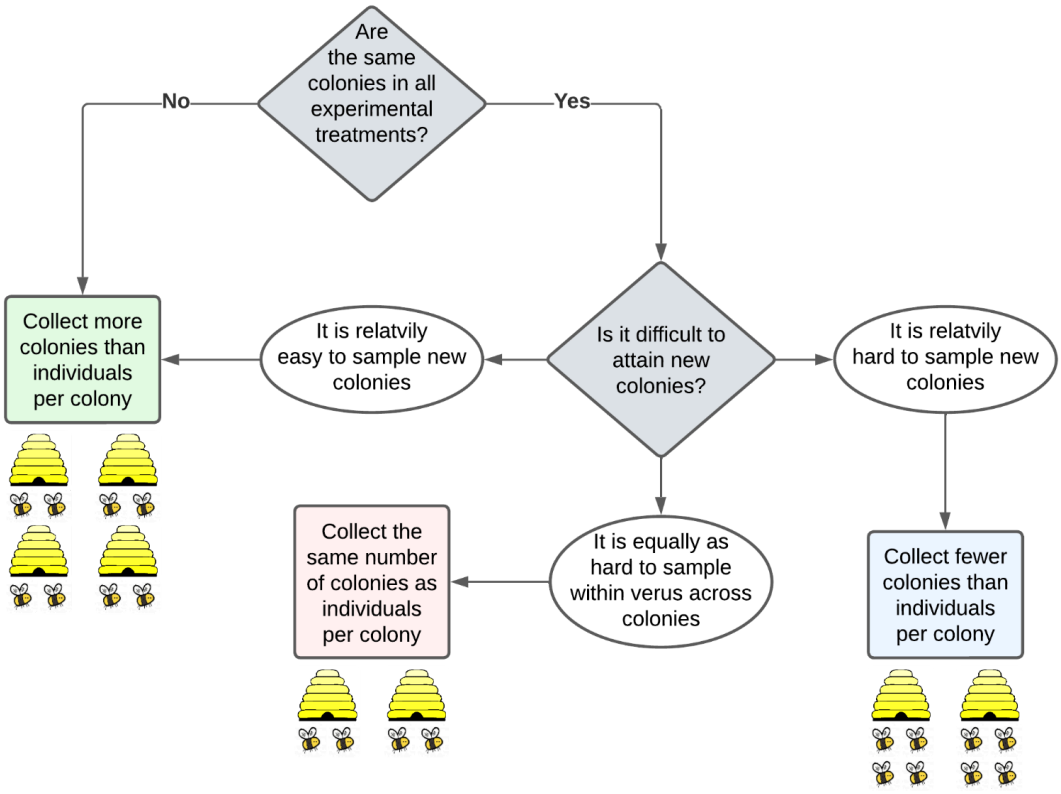


**Figure 5.** Flow diagram illustrating recommendations for experimental designs with a random effect (such as colony ID, represented as hives here). When the same colonies are present in both treatments, the design is considered a crossed design, otherwise it is a nested design.

While our sampling methodology was designed for social insect datasets structured by colonies where response variables were assumed to be normally distributed, the same sampling methodologies can be applied to many other mixed-models since LMMs are robust to even severe departures from normality (Schielzeth et al., 2020). This method can also be used for datasets with other nested structures, as in the biological sciences random effects could reflect many other scales of observation such as the year a sample was taken (Vollset et al., 2023), repeated measurements of individuals (Kentie et al., 2023), phylogenetic similarities (Dewar et al., 2024), matrilines (Power et al., 2023), and the observer (Pan & Wetzel, 2024). These results need not only apply to empiricists, as theorists could also benefit from optimal experimental design. The limiting factor of many simulations of dynamical systems is computation time, so the programmer often must trade off how long a simulation should last versus running many simulations. Running a simulation for an extended period of time is the equivalent of taking many measurements of an individual colony, maximizing precision for that single measurement. Conversely, running many simulations with different initial conditions is akin to taking measurements from different colonies. Thus, theorists may also benefit from similar sampling frameworks.

Despite the utility of a well designed experiment, power analyses tend to be underutilized in fields such as behavioral ecology and psychology (Jennions & Møller, 2003; Kiernan & Baiocchi, 2022). One potential reason for this underutilization is that it can be difficult to achieve the necessary sample sizes for real experiments. Additionally, it could be that power analyses are irrelevant to many researchers as they ignore complications imposed by random effects (Dell et al., 2022). By expanding on the definition of a power analysis to include the difficulty of sampling among groups, our study will make it easier to perform pre-experiment analyses and can therefore increase our confidence in the experimental results and avoid non-repeatable outcomes (Bak-Coleman et al., 2022).

**References**

Aarts E, Dolan CV, Verhage M, van der Sluis S. Multilevel analysis quantifies variation in the experimental effect while optimizing power and preventing false positives. BMC Neuroscience. 2015;16(1):1-15. https://doi.org/10.1186/s12868-015-0228-5

Aberson CL. Applied Power Analysis for the Behavioral Sciences. 2nd ed. New York: Routledge; 2019. https://doi.org/10.4324/9781315171500

Adams DC, Collyer ML. Multivariate phylogenetic comparative methods: evaluations, comparisons, and recommendations. Systematic Biology. 2018;67:14-31. https://doi.org/10.1093/sysbio/syx055

Aguero CM, Eyer PA, Vargo EL. Increased genetic diversity from colony merging in termites does not improve survival against a fungal pathogen. Scientific Reports. 2020;10:1-9. https://doi.org/10.1038/s41598-020-61278-7

Alpaydin E. Introduction to machine learning. MIT press; 2020.

Bak-Coleman JB, Mann RP, Bergstrom CT, Gross K, West J. Replication, varying effects, and the reliability of the scientific literature. SocArXiv. April 28, 2022.

Baker DH, Vilidaite G, Lygo FA, Smith AK, Flack TR, Gouws AD, Andrews TJ. Power contours: Optimising sample size and precision in experimental psychology and human neuroscience. Psychological Methods. 2021;26(3):295. http://dx.doi.org/10.1037/met0000337

Bengston SE, Jandt JM. The development of collective personality: the ontogenetic drivers of behavioral variation across groups. Frontiers in Ecology and Evolution. 2014;2:81. https://doi.org/10.3389/fevo.2014.00081

Bolker BM. Ecological Models and Data in R. Princeton University Press; 2008.

Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MHH, White JSS. Generalized linear mixed models: a practical guide for ecology and evolution. Trends in ecology & evolution. 2009;24:127-135. https://doi.org/10.1016/j.tree.2008.10.008

Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010, August). The balanced accuracy and its posterior distribution. In *2010 20th international conference on pattern recognition* (pp. 3121-3124). IEEE.

Bruna EM, Lapola DM, Vasconcelos HL. Interspecific variation in the defensive responses of obligate plant-ants: experimental tests and consequences for herbivory. Oecologia. 2004;138:558-565. https://doi.org/10.1007/s00442-003-1455-5

Carere C, Audebrand C, Rödel HG, d'Ettorre P. Individual behavioural type and group performance in Formica fusca ants. Behavioural processes. 2018;157:402-407. https://doi.org/10.1016/j.beproc.2018.07.009

Chaves LF. An entomologist guide to demystify pseudoreplication: data analysis of field studies with design constraints. Journal of medical entomology. 2010;47:291-298. https://doi.org/10.1093/jmedent/47.1.291

Chen G, Pine DS, Brotman MA, Smith AR, Cox RW, Taylor PA, Haller SP. Hyperbolic trade-off: the importance of balancing trial and subject sample sizes in neuroimaging. NeuroImage. 2021;118786. https://doi.org/10.1016/j.neuroimage.2021.118786

Chism G, Faron W, Davidowitz G, Dornhaus A. The influence of nest architecture on colony organization in the ant Temnothorax rugatulus. Integrative and Comparative Biology. 2019;59:37.

Cochran WG, Cox GM. Experimental Designs. New York: Wiley; 1964.

Cohen J. A power primer. In: Kazdin AE, ed. Methodological issues and strategies in clinical research. American Psychological Association; 2016. p. 279–284. https://doi.org/10.1037/14805-018

Cohen J. Statistical power analysis for the behavioral sciences. Routledge; 2013. https://doi.org/10.4324/9780203771587

Colquhoun D. An investigation of the false discovery rate and the misinterpretation of p-values. Royal Society Open Science. 2014;1(3):140216. https://doi.org/10.1098/rsos.140216

Couvillon MJ, Jandt JM, Duong NHI, Dornhaus A. Ontogeny of worker body size distribution in bumble bee (Bombus impatiens) colonies. Ecological Entomology. 2010;35:424-435. https://doi.org/10.1111/j.1365-2311.2010.01198.x

Crawley MJ. The R Book. Chichester, UK: John Wiley & Sons; 2007.

DeHeer CJ, Vargo EL. An indirect test of inbreeding depression in the termites Reticulitermes flavipes and Reticulitermes virginicus. Behavioral Ecology and Sociobiology. 2006;59:753-761. https://doi.org/10.1007/s00265-005-0105-9

Dell, R. B., Holleran, S., & Ramakrishnan, R. (2002). Sample size determination. *ILAR journal*, *43*(4), 207-213. https://doi.org/10.1093/ilar.43.4.207

Devane, D., Begley, C. M., & Clarke, M. (2004). How many do I need? Basic principles of sample size estimation. Journal of advanced nursing, 47(3), 297-302. https://doi.org/10.1111/j.1365-2648.2004.03093.x

Dewar, A. E., Belcher, L. J., Scott, T. W., & West, S. A. (2024). Genes for cooperation are not more likely to be carried by plasmids. *Proceedings of the Royal Society B*, *291*(2017), 20232549.

Fjerdingstad, E. J., Gertsch, P. J., & Keller, L. (2003). The relationship between multiple mating by queens, within‐colony genetic variability and fitness in the ant Lasius niger. *Journal of evolutionary biology*, *16*(5), 844-853.

Frank ET, Wehrhahn M, Linsenmair KE. Wound treatment and selective help in a termite-hunting ant. Proceedings of the Royal Society B: Biological Sciences. 2018;285:20172457. https://doi.org/10.1098/rspb.2017.2457

Frank MC, Braginsky M, Cachia J, Coles N, Hardwicke T, Hawkins R, Mathur B, Williams R. Experimentology: an open science approach to experimental psychology methods. https://doi.org/10.7551/mitpress/14810.001.0001

Green P, MacLeod CJ. SIMR: an R package for power analysis of generalized linear mixed models by simulation. Methods in Ecology and Evolution. 2016;7:493-498. doi: 10.1111/2041-210X.12504

Gundersen HJ, Østerby R. Optimizing sampling efficiency of stereological studies in biology: or 'Do more less well!'. Journal of microscopy. 1981 Jan;121(1):65-73. https://doi.org/10.1111/j.1365-2818.1981.tb01199.x

Harrison XA. Using observation-level random effects to model overdispersion in count data in ecology and evolution. PeerJ. 2014 Oct 9;2:e616. https://doi.org/10.7717/peerj.616

Hölldobler B, Wilson EO. The superorganism: the beauty, elegance, and strangeness of insect societies. WW Norton & Company; 2009.

Jandt JM, Bengston S, Pinter‐Wollman N, Pruitt JN, Raine NE, Dornhaus A, Sih A. Behavioural syndromes and social insects: personality at multiple levels. Biological Reviews. 2014;89:48-67. https://doi.org/10.1111/brv.12042

Jennions MD, Møller AP. A survey of the statistical power of research in behavioral ecology and animal behavior. Behavioral Ecology. 2003;14(3):438-445. https://doi.org/10.1093/beheco/14.3.438

Johnson PCD, Barry SJE, Ferguson HM, Müller P. Power analysis for generalized linear mixed models in ecology and evolution. Methods in Ecology and Evolution. 2015;6:133– 142. https://doi.org/10.1111/2041-210X.12306

Kain MP, Bolker BM, McCoy MW. A practical guide and power analysis for GLMMs: detecting among treatment variation in random effects. PeerJ. 2015;3:e1226. https://doi.org/10.7717/peerj.1226

Kentie R, Morgan Brown J, Camphuysen KC, Shamoun-Baranes J. Distance doesn't matter: migration strategy in a seabird has no effect on survival or reproduction. Proceedings of the Royal Society B. 2023 Apr 26;290(1997):20222408.

Kiernan M, Baiocchi MT. Casting new light on statistical power: an illuminating analogy and strategies to avoid underpowered trials. American Journal of Epidemiology. 2022;191(8):1500-1507. https://doi.org/10.1093/aje/kwac019

Lazic SE. Four simple ways to increase power without increasing the sample size. Laboratory animals. 2018;52(6):621-629. https://doi.org/10.1177/00236772187674

London KB, Jeanne RL. Effects of colony size and stage of development on defense response by the swarm-founding wasp Polybia occidentalis. Behavioral Ecology and Sociobiology. 2003;54:539-546. https://doi.org/10.1007/s00265-003-0662-8

Lowry CA, Montgomery DC. A review of multivariate control charts. IIE transactions. 1995;27(6):800-810. https://doi.org/10.1080/07408179508936797

Luke SG. Evaluating significance in linear mixed-effects models in R. Behavior research methods. 2017;49(4):1494-1502. https://doi.org/10.3758/s13428-016-0809-y

Pan, V. S., & Wetzel, W. C. (2024). Neutrality in plant–herbivore interactions. *Proceedings of the Royal Society B*, *291*(2017), 20232687.

Parr CL, Andersen AN, Chastagnol C, Duffaud C. Savanna fires increase rates and distances of seed dispersal by ants. Oecologia. 2007;151(1):33-41. https://doi.org/10.1007/s00442-006-0570-5

Porter EE, Hawkins BA. Latitudinal gradients in colony size for social insects: termites and ants show different patterns. The American Naturalist. 2001;157:97-106. https://doi.org/10.1086/317006

Power ML, Ransome RD, Riquier S, Romaine L, Jones G, Teeling EC. Hibernation telomere dynamics in a shifting climate: insights from wild greater horseshoe bats. Proceedings of the Royal Society B. 2023 Oct 11;290(2008):20231589.

Marshall DJ. Principles of experimental design for ecology and evolution. Ecology Letters. 2024 Apr 1;27(4):e14400-. https://doi.org/10.1111/ele.14400

Molet M, Péronnet R, Couette S, Canovas C, Doums C. Effect of temperature and social environment on worker size in the ant Temnothorax nylanderi. Journal of thermal biology. 2017;67:22-29. https://doi.org/10.1016/j.jtherbio.2017.04.013

Montgomery DC. Statistical quality control. Wiley Global Education; 2013.

Montgomery DC. Design and analysis of experiments. John Wiley & Sons; 2020.

Parr LA, Waller BM, Vick SJ. New developments in understanding emotional facial signals in chimpanzees. Current Directions in Psychological Science. 2007;16:117-122. https://doi.org/10.1111/j.1467-8721.2007.00487

Quinn GP, Keough MJ. Experimental design and data analysis for biologists. Cambridge university press; 2002 Mar 21.

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

Royston P, Sauerbrei W. Interaction of treatment with a continuous variable: simulation study of significance level for several methods of analysis. Statistics in medicine. 2013;32(22):3788-3803. https://doi.org/10.1002/sim.5813

Schielzeth H, Dingemanse NJ, Nakagawa S, Westneat DF, Allegue H, Teplitsky C, Réale D, Dochtermann NA, Garamszegi LZ, Araya-Ajoy YG. Robustness of linear mixed-effects models to violations of distributional assumptions. Methods in Ecology and Evolution. 2020;11:1141-1152. https://doi.org/10.1111/2041-210X.13434

Schulz DJ, Sullivan JP, Robinson GE. Juvenile hormone and octopamine in the regulation of division of labor in honey bee colonies. Hormones and behavior. 2002;42:222-231. https://doi.org/10.1006/hbeh.2002.1806

Tay WT, Crozier RH. Mating behaviour of Rhytidoponera sp. 12 ants inferred from microsatellite analysis. Molecular Ecology. 2001;10:167-173. https://doi.org/10.1046/j.1365-294X.2001.01167.x

Udino E, Perez M, Carere C, d'Ettorre P. Active explorers show low learning performance in a social insect. Curr Zool. 2017;63:555-560. https://doi.org/10.1093/cz/zow101

Uttley J. Power analysis, sample size, and assessment of statistical assumptions—Improving the evidential value of lighting research. Leukos. 2019 Jan 25. https://doi.org/10.1080/15502724.2018.1533851

Vargha A, Delaney HD. The Kruskal–Wallis test and stochastic homogeneity. J Educ Behav Stat. 1998;23:170-192. https://doi.org/10.3102/10769986023002170

Vollset KW, Lennox RJ, Skoglund H, Karlsen Ø, Normann ES, Wiers T, Stöger E, Barlaup BT. Direct evidence of increased natural mortality of a wild fish caused by parasite spillback from domestic conspecifics. Proceedings of the Royal Society B. 2023 Jan 25;290(1991):20221752.

Wang YA, Sparks J, Gonzales JE, Hess YD, Ledgerwood A. Using independent covariates in experimental designs: Quantifying the trade-off between power boost and Type I error inflation. J Exp Soc Psychol. 2017;72:118-124. https://doi.org/10.1016/j.jesp.2017.04.011

Wright CM, Lichtenstein JL, Doering GN, Pretorius J, Meunier J, Pruitt JN. Collective personalities: present knowledge and new frontiers. Behav Ecol Sociobiol. 2019;73:1-23. https://doi.org/10.1007/s00265-019-2639-2

**ORCID**

C.L.: 0000-0001-7238-6801
M.S: 0009-0003-0775-0170
T.P.P.: 0000-0002-7073-6932

N.M.: 0000-0002-6731-8684

**Acknowledgments**
We thank Kaitlin Baudier, Jennifer Fewell, and members of the Fewell and Pavlic labs for helpful discussion.
NM thanks Arizona State University and Okinawa Institute of Science and Technology for their support during the initial phase of this research.
This work was supported by the NSF GRFP to CL (grant number: DGE-1143953) and the JSPS Overseas Research Fellowships and JSPS Research Fellowships for Young Scientists CPD to NM (grant number: 20J00660).

**Author contributions**
C.L., M.S. N.M. conceived the idea. C.L., M.S., D.M, N.M., and T.P.P. developed theory and methodological approaches. C.L., M.S., and N.M. developed software. C.L.. M.S., and N.M. wrote the manuscript, C.L. and M.S. performed data analytics with support from N.M and T.P.P. All authors reviewed the manuscript for intellectual content and completeness and provided improvements and additional material as necessary.

**Conflict of Interest**
The authors declare no conflict of interest.