**Balancing samples within and among groups in the design of experiments from a social insect research context**

**Lynch C.M.1, Starkey M.2, Pavlic, T.P.1,2,3,4, Montgomery, D.2, Mizumoto N.1,5**

1School of Life Science, Arizona State University, Tempe, AZ, USA 85287

2School of Computing and Augmented Intelligence, Arizona State University, Tempe, AZ, USA 85287

3School of Complex Adaptive Systems, Arizona State University, Tempe, AZ, USA 85287

4School of Sustainability, Arizona State University, Tempe, AZ, USA 85287

5Okinawa Institute of Science & Technology Graduate University, Onna-son, Okinawa, Japan, 904-0495

Corresponding author: Colin Lynch

cmlynch2@asu.edu

427 E Tyler Mall, Tempe, AZ 85281

**Abstract**

1. In the fields of ecology and evolution, variance can be introduced at multiple levels in populations under study, which presents challenges for rigorous analysis. This is especially true in the social insect research context, as the individuals themselves are representatives of colonies. Colony identity is often a significant predictor of variance across individuals in a whole population, but individuals within colonies can also vary considerably. Currently, there is no framework for dividing an experiment's sampling budget across individuals or colonies.
2. The effort required to complete an experiment is derived from sample size and additional costs incurred from sampling more colonies. We compared two different sampling methods: “depth collection” that obtains more within-colony replicates than colony replicates, and “breadth collection” that obtains more colony replicates than within-colony replicates. Here, we establish a sampling budget framework by finding experimental designs which minimize the effort required in running an experiment, while also minimizing type-I (false positive) and type-II (false negative) error rates. We simulated single treatment experiments using these sampling techniques and input the resulting data into linear mixed effect models, which treat different colonies as a random effect, then use these simulations to estimate the false positive/negative rates across varying sample sizes.
3. Our literature survey estimated that sampling an additional colony is three times harder than sampling an additional replicate from within colonies. As a result of this bias, depth collection was favored over breadth collection in experiments where colonies were present in both treatments (crossed design). However, in experiments where different colonies were in each treatment (nested design), depth collection often led to high false-positive rates, and thus breadth collection outperformed depth collection despite asymmetrical sampling effort.
4. Experimental design should not only consider statistical power, but should also incorporate variable false-positive rates, asymmetrical sampling efforts, and other constraints in order to avoid sub-optimal designs and/or misleading results. By providing a time-efficient and budget-friendly experimental roadmap for researchers, this study will contribute to the future development of social insect research as well as other fields, such as ecology and evolutionary biology.

**Keywords**

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

**1. Introduction**

Datasets in ecology and evolutionary biology often have multiple-leveled structures which can drastically affect the outcomes of measurements made at each level. Individuals, genotypes, species, time periods, and other factors can all influence the interpretation of a statistical test (Schielzeth et al., 2020; Bolker et al. 2009). Datasets from social insect research are always structured by colony, a social group usually composed of a few reproductives and their non-reproductive offspring (Hölldobler & Wilson, 2009), and thus colony members share the same genetic and environmental background. 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 make-up and developmental state (Bengston & Jandt, 2014). This colony effect imposes additional variation on studies that test the effect of an experimental manipulation or the differences between populations. Ignoring such data structures results in pseudoreplication, which can confound the effects of different factors within an experiment (Chavez, 2010). Structured datasets can be analyzed by linear mixed models (LMMs), which account for the clustering of data as random effects (Bolker et al., 2009; Carere et al., 2018; Udino et al., 2017) unlike that of classical statistical models, such as ANOVA (Grilli & Rampichini, 2015). A sufficient number of colonies need to be sampled to estimate among-colony variation as a random effect. In this study, we investigate optimal sampling strategies, asking how many colonies should be sampled from when studying social insects.

When the total number of samples is limited by budget and time, an experimenter must choose how to sample at each level. There are three broad strategies for sampling across these levels: (a) depth collection, sampling a large number of replicates from a small number of colonies; (b) breadth collection, sampling a small number of replicates from a large number of colonies; or (c) balanced collection (typically referred to as a latin square design), sampling as many replicates per colony as there are number of colonies (Fig. 1). Collecting colonies in a field and/or rearing newly established colonies to add to experimental resources is likely more costly than measuring individuals from already established nests. Thus for a given sample size, depth collection is likely cheaper in many cases than breadth or balance collection. However, an insufficient number of colonies may also introduce some unintentional bias to statistical analyses. While these three strategies may or may not be equivalent in terms of costs of sampling and/or performance of statistical tests, the advantages and disadvantages of each method has never been quantified within social insect science; though they have been explored in other fields (Baker et al., 2021).

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. The analysis of the false negative rates is called power analysis which can determine the size of the sample needed to obtain a significant result when the null hypothesis is false (Jennions & Møller, 2003). 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, recent work in psychology leverages these methods to calculate power contours across subjects and trials (Baker et al., 2021; Chen et al., 2021). However these studies make the implicit assumption that the false positive rate is constant across collection strategies, ignoring the possibility that increasing power also increases false positive rates (Wang et al., 2017). Additionally, these studies either assumed an effect size or estimated an effect size using patient data. It is not clear whether the estimated effect size is above the minimal detectable threshold for their statistical tests, though our study explicitly derives this value. Furthermore, optimal sampling is achieved by also minimizing sampling efforts. In social insect research, the effort of obtaining more colonies is not equal to that of the effort required to obtain more within-colony replicates. Since a single social insect colony can contain ~100 to ~1,000,000 individuals, depending on species (London & Jeanne, 2003; Porter & Hawkins, 2001), obtaining more within-colony replicates may be easier than obtaining more colonies. Such circumstances must be taken into account when evaluating sampling strategies of which this study addresses.

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 data structured by colony. First, we surveyed the sampling strategies used in the published literature toinvestigate the relative effort to obtain within-colony replicates and among-colony replicates. 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). In crossed designs (or equivalently full factorial designs), the same colonies are used across two levels of the fixed effect, while nested designs use different colonies in each level of the fixed effect. By demonstrating how the aforementioned sampling strategies minimize false negative and false positive rates, this study provides an overview to distributing a fixed number of replicates across and within colonies.

**2. Materials and Methods**

**2.1. Literature review**

To gage what social insect scientists choose as their sampling strategies, and to categorize how they design their experiments, we surveyed 50 articles from *Insectes Sociaux*. The articles published in this journal feature only social insect research, of which this study is based on. Our primary focus was on empirical works that investigated the effect of a categorical variable (treatment), explicitly reported their sample sizes, and were published in the latest issues (volumes 65-68) so that our findings are as recent as possible. We extracted the number of within-colony replicates per colony per treatment (*I*) and the number of colonies sampled per treatment (*m*) from each paper. When *l* and *m* were not explicitly described, we calculated them by dividing the total number of colonies or replicates by the number of treatments. When there were multiple experimental variables within a single experiment, we counted the number of treatments as the total number of combinations of those variables. It is important to mention here that sample sizes were often unequal across treatment groups and thus we accounted for this observation by taking the arithmetic mean of those samples. For example, Sakamoto et al. (2020) compared the autogrooming responses of two different honey bee species, where they sampled the same number of colonies per species (*m* = 3), but different numbers of individuals per colony (71, 142, 131, 186, 35, 148). To get the number of within-colony replicates, we simply calculated the arithmetic mean of these samples (*I* = [71 + 142 + 131 + 186 + 35 + 148] / 6 ≈ 119). We then categorized the type of collection as being either breadth (*l* < *m*), depth (*l* > *m*), or balanced (*l* = *m*). If there were multiple experiments in a paper, we studied only the first experiment listed in the methods section.

We also examined the experimental design of the paper (i.e., whether nested or crossed). In crossed designs, all colonies were exposed to all treatments. For instance, Hendriksma et al. (2020) exposed individual honey bees to nicotine and dimethoate. They had 4 colonies, and tested 20 bees per colony per treatment (*m* = 4, *I* = 20). Note the difference from Sakamoto et al. (2020), which allocates different colonies to each treatment (as one colony cannot belong to two species) and is therefore considered a nested design.

We then inferred the effort of sampling more colonies to sampling more within-colony replicates. Since *I* was not equal to *m* in any of the reviewed papers, and knowing that sampling across colonies is more resource intensive for experimenters in general, we assumed this bias to be reflective of the relative sampling effort W. For each paper, we calculated the relative sampling effort as W = *I*/*m*, where W > 1 when between-colony replicates were more difficult to increase than within-colony replicates, while 0 < W < 1 when within-colony replicates were more difficult to increase than between-colony replicates. We also measured the median W across the 50 papers as a representative relative sampling effort and estimated W for each type of social insect (wasps, termites, ants, and bees) separately.

**2.2 Data generation**

To compare the performance of various sampling strategies, we expanded the methodology of Bolker (2008). Here we simulated many datasets where there was either an effect, or no effect, of an experimental treatment. We then performed an F-test on these datasets, which was derived from a linear mixed model, then tallied the number of true positives and negatives as well as the number of false positives and negatives. Doing so derives the power and balanced accuracy for each sampling strategy, both of which will be discussed later. In this study, we simulated an experiment with a single treatment that results in a mean shift with multiple replicates from multiple colonies. We assumed that the trait of interest followed a normal distribution, as this is among the most common distributions among social insect studies (Couvillon et al., 2010; Schultz et al., 2002; Parr et al., 2007; Frank et al., 2018; Tay & Crozier, 2001; DeHeer & Vargo, 2006).

We generated data by summing values drawn from two normal distributions and a fixed effect of the interest variable (0 or Δ), to be fitted to LMM treating the categorical variable as a fixed effect and colony as a random effect:

\bold y = \Delta \bold x + \alpha + \epsilon 

\alpha \sim N(0, \sigma_a^2)

\epsilon \sim N(0, \sigma_w^2)

where x represents a dummy variable taking the value of 0 (control) or 1 (treatment), σ2a represents among-colony variation, and σ2w represents within-colony variation. The first normally distributed random value was drawn for each colony, while the second was drawn for every individual social insect. In nested design, we drew different colony-specific values for each treatment to simulate different colonies between two treatments. In crossed design, we drew the colony-specific values only once so that they were shared between treatments (Fig. S2).

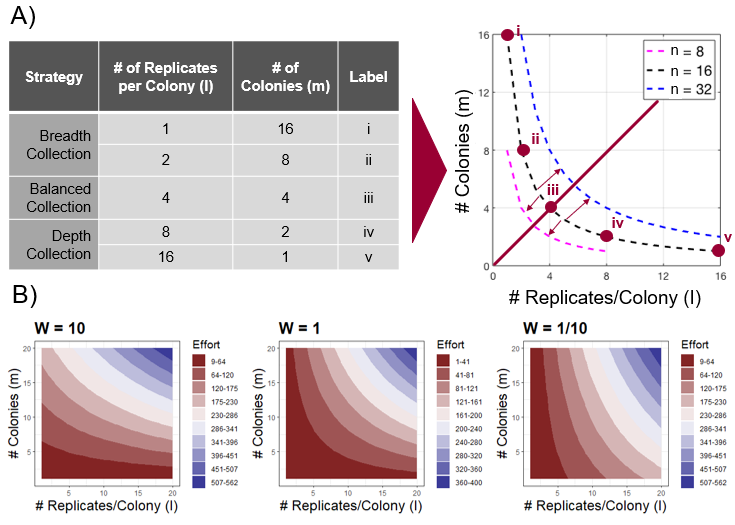
Next, we need to determine how large the sample size will be for each treatment. In each simulation, there were n replicates, with *l* replicates from *m* colonies for each treatment (n = *l* × *m*; *l, m*, n are positive integers). The sampling space is defined as the discrete region which represents all combinations of *l* and *m* (Fig. 1A). When the number of within-colony replicates is larger than the number of across-colony replicates (*l* > *m*), the sampling strategy is defined as depth collection strategy. When *l* < *m*, then it is defined as a breadth collection strategy. Finally, if *l* = *m*, then it is defined as a balanced collection strategy.

We explored all combinations of *l* and *m* in the range between 1 and 20, except (*l*, *m*) = (1, 1), for a total of 202-1 = 399 combinations. We chose an upper limit of 20 because it covers more than half of the sample sizes in the literature review (67%). For each combination of sample size, we investigated all combinations of the value of three parameters Δ, σw, and σa (the range is in Table 1, where we resampled Δ = 0 three times). In total, we have 103 = 1000 combinations of parameters. 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 performed this entire 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 2021a).

Finally, we calculate the sampling effort for each sampling strategy, which is in part derived from sample size. If the sampling cost is the same between inter- and intra-colony sampling, sampling effort is the same across different *l* and *m* combinations with a given n. In this case, sampling effort E = n = *m* × *I* (Fig. 1A), with symmetric isoeffort contours, where isoeffort contours are curves in the sampling space which represent equivalent levels of effort (Fig. 1A). However, inter and intra-colony sampling costs are rarely equal. We introduced a weight term W (which was estimated for social insects in Text 2.1) to account for the asymmetric sampling efforts:

E = mI + (W-1)m + (\frac{1}{W}-1)I.

When W > 1, then inter-colony sampling is more costly than intra-colony sampling, while when W < 1, then the opposite is true. Note that E = *m* × *I* when W = 1. With W ≠ 1, we obtain asymmetric isoeffort contours (Fig. 1B).



**Figure 1.** Required effort for different sampling strategies. (A) Mapping sampling strategies on the isoeffort contours with W = 1. When the total sample size n is fixed as 16, there are 5 possible combinations of sampling strategies (i-v). Balanced collection is on the line, *l* = *m* (iii). Breadth collections appear above the line, *l* = *m* (*l* = 1 or 2, *m* = 16 or 8, i-ii), whereas depth collections appear below it (*l* = 16 or 8, *m* = 1 or 2, iv-v). The dashed lines represent the isoeffort contours with different total sample sizes (color). (B) Isoeffort contours with differing relative difficulties (color) of sampling colonies vs sampling within colonies. When W = 10, colonies are roughly 10 times harder to sample from than individuals within a colony, 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.

|  |  |
| --- | --- |
| **Parameter** | **Range** |
| Mean Shift (Δ) | ∈ {0, 0, 0, 0.3, 0.75, 1.2, 1.65, 2.1, 2.55, 3} |
| Within Colony Variance (σ2w) | ∈ {0, (1/3)2, (2/3)2, 12, (4/3)2, (5/3)2, 22, (7/3)2, (8/3)2, 32} |
| Among Colony Variance (σ2a) | ∈ {0, (1/3)2, (2/3)2, 12, (4/3)2, (5/3)2, 22, (7/3)2, (8/3)2, 32} |

**Table 1:** Parameter values for data generation. We tested all combinations of these parameters to evaluate different sampling strategies. Note that we triplicated Δ = 0 to emphasize the importance of avoiding false positives. Small effect sizes are generally negligible (Cohen, 2016), so we set our smallest non-zero Δ at 0.3, which is the minimum detectable threshold for this test (see Text 2.4).

**2.3 Evaluation of sampling strategies**

To evaluate the performance of different sampling strategies, we used a simple LMM for each sampled dataset, where treatment was treated as the fixed effect, and colony was treated as random effect (random intercept). In crossed designs, the LMM took the form Measurement ~ Treatment + (1|Colony), but for nested designs, colony ID was nested within treatments: Measurement ~ Treatment + (1|Treatment:Colony). The statistical significance of treatment was tested by F-test. We determined whether the model could obtain the correct conclusion with a given sampling strategy with the following ruleset; when Δ > 0, true positive with p < 0.05 and estimated Δ (\hat \Delta) > 0, otherwise a false negative (type-II error); when Δ = 0, true negative with p > 0.05, otherwise false positive (type-I error). We calculated the probability of obtaining a true positive (power) and true negative, respectively. We also obtained the balanced accuracy as (true positive probability + true negative probability) / 2, to equally prioritize the minimization of both type-I and type-II errors. We evaluated how accurately the model estimated the parameters, Δ, σw, and σa by calculating squared error, e.g., (\hat{\Delta} - \Delta)^2. In addition, we fitted t-distributions and normal distributions to the sampled datasets (see Text S1.1). The t-distribution has heavier tails than a normal distribution, so if the distributions of both treatments were approximated by t-distributions, it would be less likely to estimate the true mean difference.

Using these performance metrics and effort, we can establish a framework for selecting optimal sampling strategies. The performance of sampling strategies differed depending on the combination of *l* and *m* even with a fixed sample size (Fig. 2), and sampling effort will almost always be biased (Fig. 1B, Table 2). By integrating these asymmetries, we investigated the combination of *l* and *m* values which (i) maximizes performance given a fixed amount of effort E or (ii) minimizes E given a minimum acceptable performance. As for performance, we used both power and balanced accuracy. As a case study, we explored sampling space (l < 21 and m < 21), with parameters Δ = 0 or 2.55 and σw, σa = 2 (although see Text S 1.8 for distributions of optimal strategies across parameter combinations). E was calculated with W = 2.94, a value for social insect research (Table 2). In the first approach, we set the maximum acceptable effort to 100. In the second approach, we set the minimum acceptable power to 0.8 (Crawley, 2007), while the minimum acceptable balanced accuracy was 0.9. This value of 0.9 for balanced accuracy roughly corresponds to 0.8 for power. This happens because the probability of achieving a true negative saturates to 0.975 (= 1-0.05/2) with a large enough sample size (Fig. 2E), and thus balanced accuracy (average of true negative and power) saturates to (0.8+0.975)/2 = 0.8875. Additionally, the linear regression between power and balanced accuracy found that a power of 0.8 corresponds to a balanced accuracy of 0.884 (p < 0.001, slope = 0.468, adj. R2 = 0.8287). We rounded this to 0.9 as a threshold.

Generally, we can convert power thresholds (P) to balanced accuracy thresholds (B) at a certain significance level (𝛂) for a two-sided test with the following expression:

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

While the expression for one-sided tests is:

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

All statistical tests were performed in R (R Core Team, 2021) using the stats, dunn.test, dplyr, and TTR packages. Graphs were made using ggplot2 and RColorBrewer packages.

**2.4 Minimal detectable threshold for the mean shift**

Any mean shift greater than 0 could be considered a ‘true’ mean shift in theory. However, extremely small mean shifts are unlikely to be detected by any statistical test. Here, we find the smallest detectable non-zero mean shift (Δmin) that can be detected by a cumulative sum (CUSUM) control chart. CUSUM charts are sequential analyses used to monitor small changes in a process such as manufacturing (Montgomery, 2013). CUSUM charts plot the cumulative sums of deviations of sample values from some target value. When these cumulative sums (Ci for sample i) are in-control (the true mean is the same as the target mean), Ci is a random walk around zero. However, if the process is out-of-control (the two means are different), then the cumulative sums will drift upwards or downwards. In other words, if a statistic derived from Ci (the upper control limit, C+i) exceeds a decision threshold (H), the process is considered out-of-control. Here we employed the CUSUM chart to detect a minimal meaningful mean shift between the control and treatment groups (Δmin). The variations were set at the maximum value within our simulations (σw = σa = 9), and the sample size per treatment at its highest (n = *Im* = 202 = 400). As the variance of two summed normal distributions is the sum of the variances, the standard deviation of the process (σ) is \sqrt{\sigma^2_a+\sigma^2_w}.

Let µ0 be the mean of the control treatment, µ1 is the mean of the treatment group; µ0 = 0 and µ1 = µ0 + Δmin = Δmin. The cumulative sum of the observation, x, at sample i is:

C_i = \sum_{j=1}^i(x_j-\mu_0) = \sum_{j=1}^ix_j

For samples in the control group, Ci will hover around 0. However, in the treatment groups, Ci will increase on average by a value of Δmin for every sequential sample. We want to know if Ci will cross H at the last possible sample at i = n = 400. We do this by first calculating the one-sided upper CUSUM C+i, which is given by the expression:

C_i^+ = \text{max}[0, x_i-(\mu_0 +K)+C_{i-1}^+]

where K is the slack value and is usually halfway between the target mean and the out-of-control mean, so |µ1 - µ0|/2 = Δmin/2 (Montgomery, 2013). K can be understood as the smallest amount of shift we allow a process to undergo before consecutive deviations to start accumulating. If K = 0, then the CUSUM chart could experience a high number of false positives.

In the treatment group, xi is drawn from the sum of two normal distributions whose means are 0 and Δmin. The mean of the resulting joint distribution is the sum of the means, so the expected value of xi is Δmin, hence:

C_i^+ = \text{max}[0, \Delta_{\text{min}} - \frac{\Delta_{\text{min}}}{2} + C_{i-1}^+] = \text{max}[0, \frac{\Delta_{\text{min}}}{2} + C_{i-1}^+] 

As we are searching for a positive value of Δmin and C+i is initialized to 0 when i = 0, then Δmin + C+i-1 > 0. We can therefore ignore the max function:

C_i^+ = \frac{\Delta_{\text{min}}}{2} + C_{i-1}^+

Δmin/2 is a constant, so we can expand this expression and simplify:

C_i^+ = 0 + \frac{\Delta_{\text{min}}}{2} + \frac{\Delta_{\text{min}}}{2} + ...+ \frac{\Delta_{\text{min}}}{2} = i\frac{\Delta_{\text{min}}}{2}

The decision interval H is usually taken to be 5σ (Montgomery, 2013). Δmin will allow us to detect an out-of-control process with the last possible sample, so i = n. We set H equal to C+i :

n \frac{\Delta_{\text{min}}{2} = 5\sqrt{\sigma^2_a+\sigma^2_w} 

We can then finally solve for Δmin, whose value is approximately 0.3:

\Delta_{\text{min}} = \frac{10\sqrt{\sigma^2_a+\sigma^2_w} }{n} \approx 0.3

**3. Results**

**3.1 Sample size used in social insect research**

We found that most papers used a depth collection rather than a breadth collection (35 out of 50 used depth collection, binomial test p < 0.01). Therefore, we estimated that sampling more colonies was more costly than sampling more from within colonies. Therefore, the relative effort of sampling more colonies to sampling more within-colony replicates was higher than 1 (W = 2.94). The estimated W was variable across social insect taxa, but all of them had W larger than 1 (Table 2). Consistently, studies sampled more within-colony replicates than the number of colonies (Wilcoxon test W = 1543.5, p < 0.05, median within colony samples *=* 19, median across colony samples = 9). This trend held for both nested and crossed designs (Fig. S1). Although studies that used breadth collection sampled more coloniesthan depth collection, the number of within colony samplesfor depth collections were proportionally higher than the number of colonies for breadth collection (Fig. S1). The median within and across colony sample sizesfor depth collection was 22 and 4 respectively, while that for breadth collection was 4 and 14.

Crossed designs were used as often as nested designs (23 out of 50 used crossed designs, binomial test p = 0.6718). 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 W = 352, p = 0.3555).

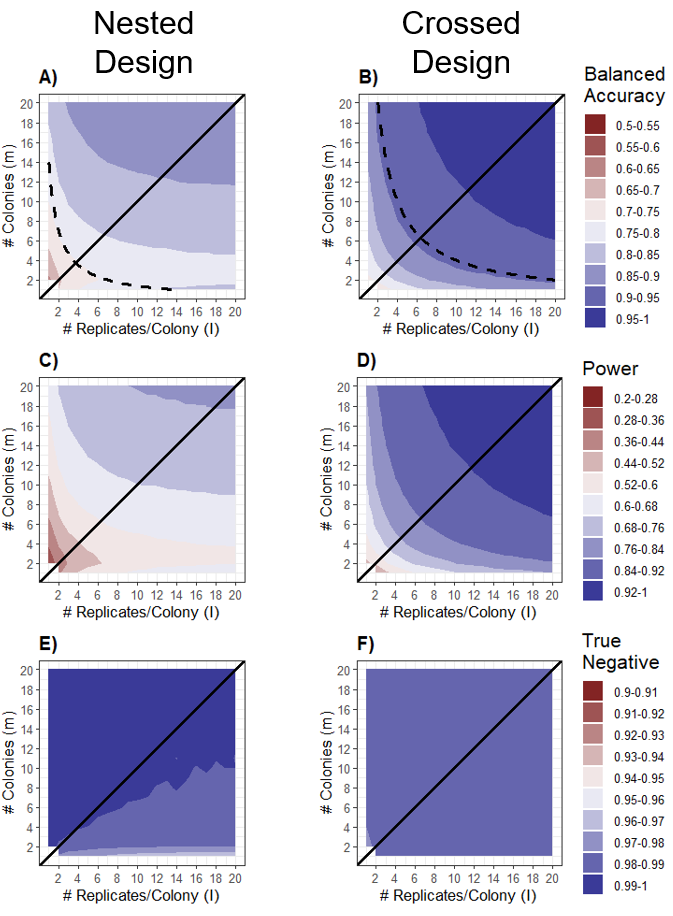
|  |  |
| --- | --- |
|  | **Median W (sample size)** |
| Ant | 2.92 (19) |
| Bee | 5 (12) |
| Wasp | 1.14 (10) |
| Termite | 21.4 (9) |
| **Overall** | ***2.94 (50)*** |

**Table 2:** Estimates of W from literature review.

**3.2 Relative performance of sampling strategies**

For crossed designs, balanced accuracy and power did not depend on the sampling strategy. Rather, it depended purely on sample size (Fig. 2 BDF). Put another way, isoeffort curves where W = 1 (E = n = *mI*) follow the contour plots of balanced accuracy and power in the sampling space (Fig. 2B). Additionally, the probability of a true negative (or 1 - probability of a false positive) is constant regardless of sample size or sampling strategy (Fig. 2F). Crossed designs also tended to have higher power, balanced accuracy, and true negative probabilities than nested designs (Fig. 2).

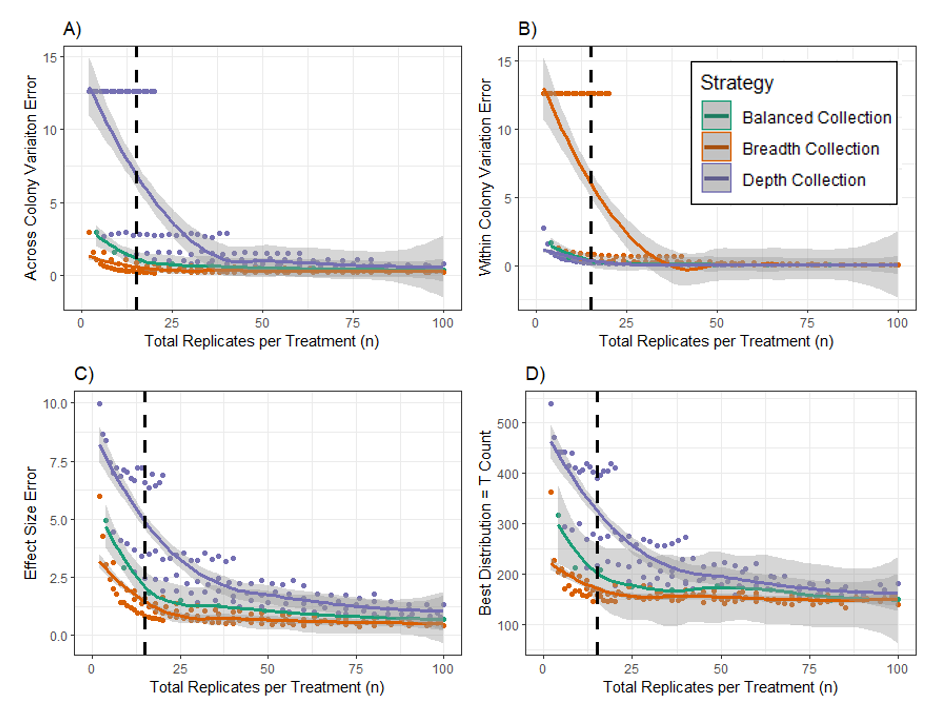
On the other hand, the performance of sampling strategies in nested designs depends on the sampling strategy (Fig. 2ACE). Breadth collection generally achieved higher balanced accuracy than depth collection. Namely, balanced accuracy increased with more among-colony replicates, rather than more within-colony replicates (Fig. 2A). The exception was at very small sample sizes (n < 15), where depth collections resulted in a higher balanced accuracy (Fig. 2A, Text S1.3). In this region, depth collection was sensitive to even small differences between treatments, leading to higher power as well as lower true negative rates (Fig. 3CE). With this trade-off, balanced accuracy preferred depth collection because depth collection improved power to a higher degree than breadth collection improved true negatives. Thus, balanced accuracy may have overestimated the utility of depth collection at small sample sizes. With a large enough sample size (n > 15), breadth collections achieved both higher power and a higher true negative probability than depth collection (Fig. 2CE).



**Figure 2.** The relative advantage of depth and breadth collection across different sample sizes pooled across all parameter combinations. The performance was evaluated by either power (C, D), true negative probability (E, F), or balanced accuracy (A, B). Nested designs are in the first column (A, C, E), while crossed designs are in the second column (B, D, F). The solid diagonal lines represent balanced collection strategies (*l* = *m*). Above the line shows breadth collection strategies (*m* > *I*) while depth collection strategies lie below it (*m* < *I*). The dotted line in A is the isoeffort curve (n = 14, W = 1). Below this line, depth collection tends to have a higher balanced accuracy than breadth collection. A similar threshold but far less prominent at n = 40 exists for crossed design (B).

In nested 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. 3A), while depth collection estimated within-colony variance better than breadth collections (Fig. 3B). 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. 3AB). 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. When the true σw and σa values were the same, breadth collection estimated σw better than depth collection estimated σa (one-sample one-sided Wilcoxon signed rank test: V = 4,338,255, p < 0.001). This means that overall breadth collection had a better estimate of total variance than depth collection.

Breadth collection also minimized effect size error irrespective of sample sizes (Fig. 3C). This happened because datasets produced by breadth collection robustly generated normal distribution even at small sample size. Conversely, datasets produced by depth collection often resulted in t-distribution-like distributions (Text S 1.1), rather than normal distributions, especially at small sample size (Fig. 3D). There was a significant positive correlation between t-distribution count and effect size error (linear regression p < 0.001, R2 = 0.9756).

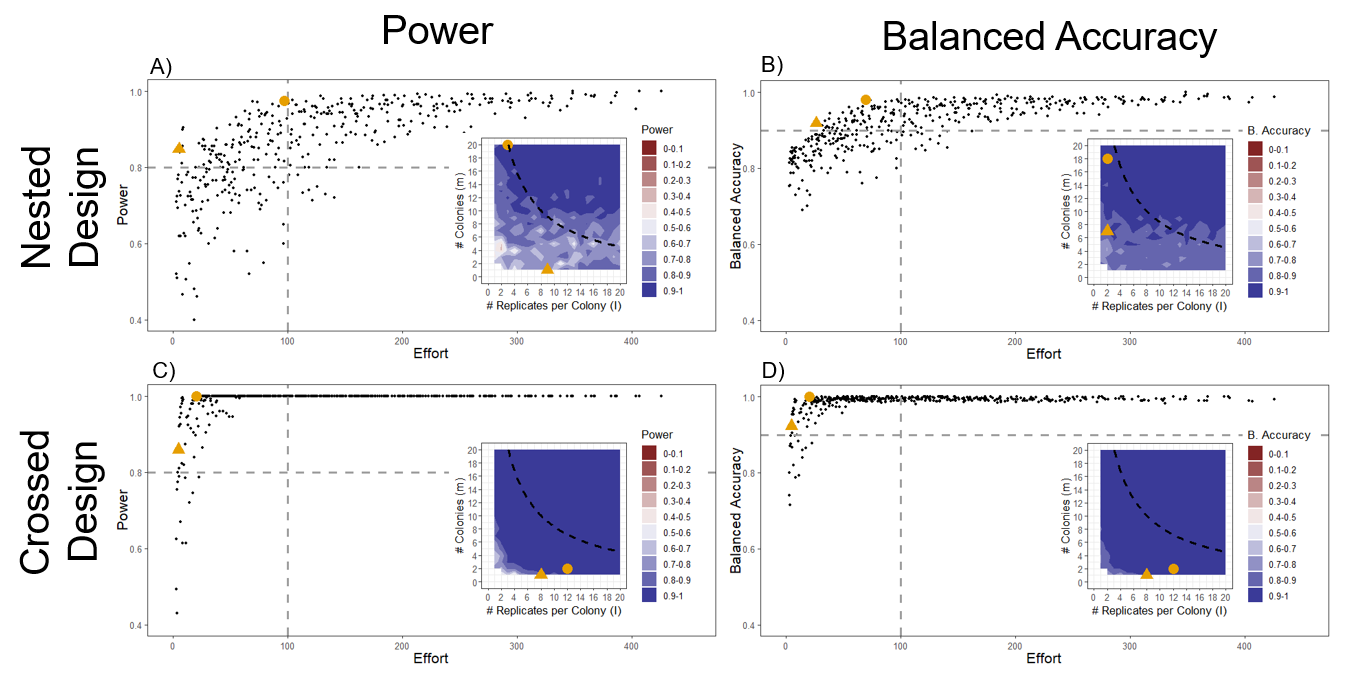


**Figure 3.** The performance of sampling strategies across different numbers of replicates per treatment (n) for nested designs. The solid colored lines show the LOESS regression for each collection type across n, and the shaded regions indicate 95% confidence intervals. The dotted black line is n = 14, best separation of breadth and depth collection in terms of balanced accuracy (Fig. 2A). The x-axis is limited to n = 100 for visibility.

**3.3 Sampling strategies accounting for sampling efforts case study**

In social insect research, depth collection required less effort than breadth collection at a fixed sample size (W > 1, Table 2). Therefore, in crossed designs, depth collection is favored over breadth collection as both collection methods were equivalent in the performance (Fig. 2BDF). In this example (Δ = 0 or 2.55 and σw, σa = 2), depth correction was preferred over breadth collection after accounting for the asymmetric sampling efforts (W = 2.94), where (*I, m*) = (12, 2) maximized power when the upper limit of E was 100. Alternatively, (*I, m*) = (8, 1) minimized E to achieve power of at least 0.8 (Fig. 4C). The results are identical for balanced accuracy, where (12, 2) maximized balanced accuracy when E had an upper limit at 100. (8, 1) minimized E to achieve a balanced accuracy of 0.9.

In nested designs, breadth collection was preferred over the depth collection despite W > 1. When we focused on power, the results of each optimization method were contradictory, where breadth collection, (*I, m*) = (2, 20), maximized power when the effort was limited to 100, while depth collection while (9, 1), minimized E to achieve a power of 0.8. However, this is problematic because the strategy (9, 1) fell into the region where the false positive probability was higher than the conventional threshold, 0.05 (Fig. 2B). This did not occur when we used balanced accuracy as the performance metric. Here, the optimal strategy was (2, 7) rather than (9, 1), showing that breadth collection can help us avoid regions with high false-positive rates (Fig. 4B). The optimal strategy for maximizing balanced accuracy given the E limit was similar to that of the power estimate at (2, 18). These patterns emerged across different parameter combinations (Fig. S8).



**Figure 4.** Optimal sampling strategies after accounting for effort. Each external plot shows the relationship between effort and power (A, C) or balanced accuracy (B, D) with parameters W = 2.94, Δ = 0 or 2.55 and σw, σa = 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 gold circle is the strategy that maximizes power given an upper limit on effort. The gold triangle is the strategy that minimizes effort given a lower limit on power or balanced accuracy. These optimal strategies correspond to the gold circle and triangle on the inset plots, which show the sampling space for each design. The black curved lines on the contour plot correspond to the dashed vertical lines.

**4. Discussion**

Our results highlight the gap between ideal sampling strategies and real experimental practice used in social insect research. We found that previous studies predominantly used depth collection rather than breadth collection, whether experiment is nested or crossed design. However, breadth collection can more effectively test experimental hypotheses in nested designs (Fig. 2). The advantage of breadth collection in nested design experiments is maintained even after accounting for the fact that sampling colonies is generally more difficult than sampling within colonies (Fig. 4). Note that the degree of difficulty depends on social insect taxa, where obtaining many colonies is easier for species with open nests and small colony size like wasps, while it is more difficult for species with cryptic lifestyles and large colony sizes, like termites. Additionally, although the sample size differed by a few orders of magnitude (6 to ≈ 47,000) among studies, sample size was not different between nested and crossed designs. This contradicts proper experimental practice, where nested designs require larger sample sizes as they cannot control sources of variation from within the nested structure (Lazic, 2018). In summary, social insect researchers can profit from sampling additional colonies when they need to use nested experimental designs. For crossed designs, depth collection is sufficient when the experimental goal is to detect an effect of treatment.

Breadth collection is preferred over depth collection in nested designs because depth collection can lead to high false-positive probability at small sample sizes. For n < 15, 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 is the mean of true positive and true negative probability, where scientists accept false negatives (typically β = 0.2) four times as easily as false positive (typically α = 0.05), so at these thresholds balanced accuracy is not sensitive to small deviations in the false positive rate. These false positives arise because sample distributions have fatter than expected tails, resembling t-distributions rather than normal distributions. Thus, the difference in sample means between control and treatment groups can be inflated, resulting in not only magnifying a true signal (true positive probability 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 tends to outweigh 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. 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 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 high.

However, this region tends to be selected more often when power is used as the performance metric (Fig 4A; Fig S8A). This occurs because power analyses rely on solely type-II error (false negative) rates to give estimates of sample size (Cohen, 2013) because it is generally assumed that type-I error (false positive) rate is consistent for most experimental setups given the conventional threshold, 0.05 (Aberson, 2019; Royston & Sauerbrei, 2013). This implies that power analyses cannot be used in cases where the type-I error rates can be variable, e.g., with small sample sizes (Luke, 2017) with multiple comparisons (Colquhoun, 2014), or when nested data structures are ignored (Aarts et al., 2015). We show that power analysis can be problematic when the total replicates are limited by experimental constraints. In nested designs, using only power can result in a suboptimal design because depth collections not only obtain lower type-II error rates but also higher type-I error rates, compared to breadth collections (Fig. 4). On the other hand, power is a sufficient metric for crossed designs as the false-positive rate is constant across sample sizes and sampling strategies. Generally, we recommend considering both type-I and type-II errors to determine the sample size when the sampling effort is limited. For example, we used balanced accuracy to equally weight both types of error. However, if a particular study needs to minimize one type of error more than another, then one could use a weighted average between power and the true negative rate instead.

We also expanded the notion of a sample size as well by considering asymmetric sampling efforts for within-colony and among-colony replicates. Our literature review indicates that it is more difficult to sample across colonies rather than within colonies. Performance improves with larger sample sizes, but this also increases overall effort. Thus, optimization can be achieved by maximizing performance given a maximum amount of effort, or minimizing effort given a required performance. Overall, breadth collection outperformed depth collection for nested designs, while depth collection outperformed breadth collection in crossed designs. We believe that this general practice is valid for most social insect researchers because the results are held across different parameter combinations (S 1.5, S 1.6, S 1.7). Furthermore, as LMMs are robust to even severe departures from model assumptions (Schielzeth et al. 2020), these patterns could possibly even hold for non-normal error distributions. If nothing is known a priori about the system, then we estimate that the optimal solution for nested designs is to collect 4 replicates from 10 colonies (Fig. S8). Still, sampling strategies should be tuned for individual research. We provide code that automatically calculates optimal strategies for a simple two-treatment design, but this work can be expanded in the future to encompass other more complicated designs.

While our sampling methodology was designed for social insect datasets structured by colonies, the same sampling methodologies can be applied to any other random effect of an LMM or GLMM. These random effects can include the looms of a textile company (Montgomery, 2020), field sites (Hector et al., 2010), repeated measurements from individuals (Proust & Jacqmin-Gadda, 2005), phylogenetic clades (Adams & Collyer, 2018), collections of functional traits in botany (Messier et al., 2010), blocks (Masciocchi, 2013), demographics (Engemann, 2009), as well as genotypes and species (Bolker et al., 2009). Power analyses tend to be underutilized in fields such as behavioral ecology (Jennions & Møller, 2003), and one potential reason for this reluctance is that it is sometimes very difficult to achieve the necessary sample sizes for real experiments. By expanding on the definition of a power analysis to include 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.

**Literature Review References**

Alma, A. M., Farji-Brener, A. G., & Elizalde, L. (2019). When and how obstacle size and the number of foragers affect clearing a foraging trail in leaf-cutting ants. *Insectes Sociaux*, **66**, 305-316. https://doi.org/10.1007/s00040-018-00680-x

Birch, G., Cant, M. A., & Thompson, F. J. (2019). Behavioural response of workers to repeated intergroup encounters in the harvester ant *Messor barbarus*. *Insectes Sociaux*, **66**, 491-500. https://doi.org/10.1007/s00040-019-00710-2

Bordoni, A., G. Mocilnik, G. Forni, M. Bercigli, C. D. V. Giove, A. Luchetti, S. Turillazzi, L. Dapporto, and M. Marconi.(2020). Two aggressive neighbours living peacefully: the nesting association between a stingless bee and the bullet ant. *Insectes Sociaux*, **67**, 103-112. https://doi.org/10.1007/s00040-019-00733-9

Branconi, R., Baracchi, D., Turillazzi, S., & Cervo, R. (2018). Testing the signal value of clypeal black patterning in an Italian population of the paper wasp *Polistes dominula*. *Insectes sociaux*, **65**, 161-169. https://doi.org/10.1007/s00040-017-0598-z

Brito, E. L. S., Aragão, M., & Santos, G. M. M. (2018). Colony defensive behavior by the swarm-founding wasp *Parachartergus pseudoapicalis*: increase on investment predicts the intensity of nest defense. *Insectes Sociaux*, **65**, 411-417. https://doi.org/10.1007/s00040-018-0627-6

Brown, M., & Brown, M. J. F. (2020). Nectar preferences in male bumblebees. *Insectes Sociaux*, **67**, 221-228. https://doi.org/10.1007/s00040-020-00751-y

Chouvenc, T. (2019). The relative importance of queen and king initial weights in termite colony foundation success. *Insectes Sociaux*, **66**, 177-184. <https://doi.org/10.1007/s00040-019-00690-3>

Chouvenc, T. (2020). Limited survival strategy in starving subterranean termite colonies. *Insectes Sociaux*, **67**, 71-82. https://doi.org/10.1007/s00040-019-00729-5

Cini, A., Branconi, R., Patalano, S., Cervo, R., & Sumner, S. (2020). Behavioural and neurogenomic responses of host workers to social parasite invasion in a social insect. *Insectes Sociaux*, **67**, 295-308. https://doi.org/10.1007/s00040-020-00765-6

Costa, L., Nunes-Silva, P., Galaschi-Teixeira, J. S., Arruda, H., Veiga, J. C., Pessin, G., & Imperatriz-Fonseca, V. L. (2021). RFID-tagged amazonian stingless bees confirm that landscape configuration and nest re-establishment time affect homing ability. *Insectes Sociaux*, **68**, 101-108. https://doi.org/10.1007/s00040-020-00802-4

Csata, E., Billen, J., Bernadou, A., Heinze, J., & Markó, B. (2018). Infection-related variation in cuticle thickness in the ant *Myrmica scabrinodis* (Hymenoptera: Formicidae). *Insectes Sociaux*, **65**, 503-506. https://doi.org/10.1007/s00040-018-0628-5

Cunha, M. S., Campos, L. A. O., & Lopes, D. M. (2020). Insights into the heterochromatin evolution in the genus *Melipona* (Apidae: Meliponini). *Insectes Sociaux*, **67**, 391-398. https://doi.org/10.1007/s00040-020-00773-6

Czaczkes, T. J., & Kumar, P. (2020). Very rapid multi-odour discrimination learning in the ant *Lasius niger*. *Insectes Sociaux*, **67**, 541-545. https://doi.org/10.1007/s00040-020-00787-0

Delattre, O., J. Šobotník, V. Jandák, J. Synek, J. Cvačka, O. Jiříček, T. Bourguignon, and David Sillam-Dussès (2019). Chemical and vibratory signals used in alarm communication in the termite *Reticulitermes flavipes* (Rhinotermitidae). *Insectes Sociaux*, **66**, 265-272. https://doi.org/10.1007/s00040-018-00682-9

Derstine, N. T., Gries, R., Zhai, H., Jimenez, S. I., & Gries, G. (2018). Cuticular hydrocarbons determine sex, caste, and nest membership in each of four species of yellowjackets (Hymenoptera: Vespidae). *Insectes Sociaux*, **65**, 581-591. https://doi.org/10.1007/s00040-018-0649-0

Doering, G. N., Kamath, A., Wright, C. M., & Pruitt, J. N. (2018). Evidence for contrasting size-frequency distributions of workers patrolling vegetation vs. the ground in the polymorphic African ant Anoplolepis custodiens. *Insectes Sociaux*, **65**, 663-668. https://doi.org/10.1007/s00040-018-0645-4

Eloff, J., Veldtman, R., Bulgarella, M., & Lester, P. J. (2020). Population genetics of the invasive wasp *Vespula germanica* in South Africa. *Insectes Sociaux*, **67**, 229-238. https://doi.org/10.1007/s00040-020-00752-x

Eriksson, T. H., Hölldobler, B., Taylor, J. E., & Gadau, J. (2019). Intraspecific variation in colony founding behavior and social organization in the honey ant *Myrmecocystus mendax*. *Insectes Sociaux*, **66**, 283-297. https://doi.org/10.1007/s00040-019-00687-y

Felden, A., Baty, J. W., Bulgarella, M., Brown, R. L., Dobelmann, J., Gruber, M. A. M., & Lester, P. J. (2020). Viral and fungal pathogens associated with *Pneumolaelaps niutirani* (Acari: Laelapidae): a mite found in diseased nests of Vespula wasps. *Insectes Sociaux*, **67**, 83-93. https://doi.org/10.1007/s00040-019-00730-y

Glass, J. R., Duell, M., & Harrison, J. F. (2020). Defensive biting by *Tetragonisca angustula* is dangerous but not suicidal. *Insectes Sociaux*, **67**, 515-522. https://doi.org/10.1007/s00040-020-00790-5

Harano, K., Maia-Silva, C., & Hrncir, M. (2020). Why do stingless bees (*Melipona subnitida*) leave their nest with resin loads?. *Insectes Sociaux*, **67**, 195-200. https://doi.org/10.1007/s00040-019-00748-2

Hellemans, S., Fournier, D., Hanus, R., & Roisin, Y. (2019). Sex ratio variations among years and breeding systems in a facultatively parthenogenetic termite. *Insectes sociaux*, **66**, 129-138. https://doi.org/10.1007/s00040-018-0667-y

Hendriksma, H. P., Bain, J. A., Nguyen, N., & Nieh, J. C. (2020). Nicotine does not reduce *Nosema ceranae* infection in honey bees. *Insectes Sociaux*, **67**, 249-259. https://doi.org/10.1007/s00040-020-00758-5

Janowiecki, M., Clifton, E., Avalos, A., & Vargo, E. L. (2020). Upper thermal tolerance of tropical and temperate termite species (Isoptera: Rhinotermitidae, Termitidae): a test of the climate variability hypothesis in termites. *Insectes Sociaux*, **67**, 51-57. https://doi.org/10.1007/s00040-019-00727-7

Lash, C. L., Sturiale, S. L., Kisare, S. A., & Kwit, C. (2020). The effects of a myrmecochore-produced chemical on entomopathogenic fungal growth and seed-dispersing ant survival rates and foraging patterns. *Insectes Sociaux*, **67**, 495-505. https://doi.org/10.1007/s00040-020-00786-1

Leitner, N., Lynch, C., & Dornhaus, A. (2019). Ants in isolation: obstacles to testing worker responses to task stimuli outside of the colony context. *Insectes Sociaux*, **66**, 343-354. https://doi.org/10.1007/s00040-019-00692-1

Liu, J. F., Wan, X. J., Wang, Z. L., He, X. J., & Zeng, Z. J. (2020). Influence of RNA interference-mediated reduction of Or11 on the expression of transcription factor Kr-h1 in Apis mellifera drones. *Insectes Sociaux*, **67**, 411-418. https://doi.org/10.1007/s00040-020-00775-4

Macháčková, L., Votavová, A., Mikát, M., Matějková, S., Řehoř, I., Gillarová, S., & Straka, J. (2019). Measurement of protein and sugar consumed by bumblebee larvae under standard and food stress conditions using lanthanide complexes. *Insectes Sociaux*, **66**, 245-256. https://doi.org/10.1007/s00040-018-00681-w

Malé, P. J. G., Youngerman, E., Pierce, N. E., & Frederickson, M. E. (2020). Mating system, population genetics, and phylogeography of the devil’s garden ant, *Myrmelachista schumanni*, in the Peruvian Amazon. *Insectes Sociaux*, **67**, 113-125. https://doi.org/10.1007/s00040-019-00735-7

Melgarejo, V., Wilson Rankin, E. E., & Loope, K. J. (2018). Do queen cuticular hydrocarbons inhibit worker reproduction in *Bombus impatiens*?. *Insectes Sociaux*, **65**, 601-608. https://doi.org/10.1007/s00040-018-0651-6

Montagu, A., Lee, T. R. C., Ujvari, B., McCarl, V., Evans, T. A., & Lo, N. (2020). High numbers of unrelated reproductives in the Australian ‘higher’ termite *Nasutitermes exitiosus* (Blattodea: Termitidae). *Insectes Sociaux*, **67**, 281-294.

O’Donnell, S., Bulova, S., Caponera, V., Oxman, K., & Giladi, I. (2020). Species differ in worker body size effects on critical thermal limits in seed-harvesting desert ants (*Messor ebeninus* and *M. arenarius*). *Insectes Sociaux*, **67**, 473-479. https://doi.org/10.1007/s00040-020-00782-5

Palavalli-Nettimi, R., & Narendra, A. (2018). Does size affect orientation using celestial cues?. *Insectes Sociaux*, **65**, 657-662. https://doi.org/10.1007/s00040-018-0640-9

Ruano, F., Lenoir, A., Silvestre, M., Khalil, A., & Tinaut, A. (2019). Chemical profiles in *Iberoformica subrufa* and *Formica frontalis,* a new example of temporary host–parasite interaction. *Insectes Sociaux*, **66**, 223-233. https://doi.org/10.1007/s00040-018-00677-6

Sakamoto, Y., Maeda, T., Yoshiyama, M., Konno, F., & Pettis, J. S. (2020). Differential autogrooming response to the tracheal mite *Acarapis woodi* by the honey bees *Apis cerana* and *Apis mellifera*. *Insectes Sociaux*, **67**, 95-102. https://doi.org/10.1007/s00040-019-00732-w

Sandoval-Arango, S., Cárdenas Henao, H., & Montoya-Lerma, J. (2020). Divergence in Bergmann’s clines: elevational variation and heritability of body size in a leaf-cutting ant. *Insectes Sociaux*, **67**, 355-366.

Smith, A. A. (2019). Prey specialization and chemical mimicry between *Formica archboldi* and *Odontomachus* ants. *Insectes Sociaux*, **66**, 211-222. https://doi.org/10.1007/s00040-018-0675-y

Southon, R. J., Radford, A. N., & Sumner, S. (2020). High reproductive skew in the Neotropical paper wasp *Polistes lanio*. *Insectes Sociaux*, **67**, 451-456. https://doi.org/10.1007/s00040-020-00780-7

Tanaka, H. O., Haraguchi, T. F., Tayasu, I., & Hyodo, F. (2019). Stable and radio-isotopic signatures reveal how the feeding habits of ants respond to natural secondary succession in a cool-temperate forest. *Insectes Sociaux*, **66**, 37-46. https://doi.org/10.1007/s00040-018-0665-0

Takata, M., Inagaki, T., Ishibashi, T., Tasaki, E., & Matsuura, K. (2020). A non-invasive method for sexing first and second instar larvae of termites using external morphology. *Insectes Sociaux*, **67**, 487-493. https://doi.org/10.1007/s00040-020-00785-2

Tawdros, S., West, M., & Purcell, J. (2020). Scaling relationships in *Formica* ants with continuous worker size variation. *Insectes Sociaux*, **67**, 463-472. https://doi.org/10.1007/s00040-020-00779-0

Unnikrishnan, S., & Gadagkar, R. (2021). Dominance behaviour and division of labour in the tropical primitively eusocial wasp *Ropalidia cyathiformis*. *Insectes Sociaux*, **68**, 123-132. https://doi.org/10.1007/s00040-020-00803-3

Uy, F. M. K., Adcock, J. D., Jeffries, S. F., & Pepere, E. (2019). Intercolony distance predicts the decision to rescue or attack conspecifics in weaver ants. *Insectes Sociaux*, **66**, 185-192. https://doi.org/10.1007/s00040-018-0674-z

van Langevelde, F., Kiggen, F., van Dooremalen, C., & Cornelissen, B. (2020). Corpse removal increases when honey bee colonies experience high *Varroa destructor* infestation. *Insectes Sociaux*, **67**, 507-513. https://doi.org/10.1007/s00040-020-00789-y

Veisi, H., Nazemi-Rafie, J., Azizi, A., & Rajabi-Maham, H. (2020). Genetic characteristics of the Iranian honey bee, *Apis mellifera meda*, based on mitochondrial genes of ND4, ND4L and ND6 and their internal transcribed spacers. *Insectes Sociaux*, **67**, 439-448. https://doi.org/10.1007/s00040-020-00769-2

Welch, L. E., Baudier, K. M., & Harrison, J. F. (2020). Warmer mid-day temperatures increase leaf intake by increasing forager speed and success in *Atta colombica* during the rainy season. *Insectes Sociaux*, **67**, 213-219. https://doi.org/10.1007/s00040-020-00749-6

Woon, J. S., Boyle, M. J. W., Ewers, R. M., Chung, A., & Eggleton, P. (2019). Termite environmental tolerances are more linked to desiccation than temperature in modified tropical forests. *Insectes sociaux*, **66**, 57-64. https://doi.org/10.1007/s00040-018-0664-1

Yamanaka, Y., Iwata, R., & Kiriyama, S. (2019). Cannibalism associated with artificial wounds on the bodies of *Reticulitermes speratus* workers and soldiers (Isoptera: Rhinotermitidae). *Insectes Sociaux*, **66**, 107-117. https://doi.org/10.1007/s00040-018-0661-4

Yoshimura, H., & Yamada, Y. Y. (2018). The first brood emerges smaller, lighter, and with lower lipid stores in the paper wasp *Polistes jokahamae* (Hymenoptera: Vespidae). *Insectes Sociaux*, **65**, 473-481. https://doi.org/10.1007/s00040-018-0636-5

Yoshimura, H., & Yamada, Y. Y. (2021). Preimaginal caste-related bias in the paper wasp *Polistes jokahamae* is limited to the first brood. *Insectes Sociaux*, **68**, 133-143. https://doi.org/10.1007/s00040-020-00805-1

**References**

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

Aberson, C. L. (2019). Applied Power Analysis for the Behavioral Sciences (2nd ed). New York: Routledge. https://doi.org/10.4324/9781315171500

Adams, D. C., & Collyer, M. L. (2018). Multivariate phylogenetic comparative methods: evaluations, comparisons, and recommendations. *Systematic biology*, **67**, 14-31. https://doi.org/10.1093/sysbio/syx055

Aguero, C. M., Eyer, P. A., & Vargo, E. L. (2020). Increased genetic diversity from colony merging in termites does not improve survival against a fungal pathogen. *Scientific reports*, **10**, 1-9. https://doi.org/10.1038/s41598-020-61278-7

Baker, D. H., Vilidaite, G., Lygo, F. A., Smith, A. K., Flack, T. R., Gouws, A. D., & Andrews, T. J. (2021). Power contours: Optimising sample size and precision in experimental psychology and human neuroscience. *Psychological Methods*, *26*(3), 295. http://dx.doi.org/10.1037/met0000337

Bengston, S. E., & Jandt, J. M. (2014). The development of collective personality: the ontogenetic drivers of behavioral variation across groups. *Frontiers in Ecology and Evolution*, **2**, 81. https://doi.org/10.3389/fevo.2014.00081

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J. S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in ecology & evolution*, **24**, 127-135. https://doi.org/10.1016/j.tree.2008.10.008

Bruna, E. M., Lapola, D. M., & Vasconcelos, H. L. (2004). Interspecific variation in the defensive responses of obligate plant-ants: experimental tests and consequences for herbivory. *Oecologia*, **138**, 558-565. https://doi.org/10.1007/s00442-003-1455-5

Carere, C., Audebrand, C., Rödel, H. G., & d’Ettorre, P. (2018). Individual behavioural type and group performance in Formica fusca ants. *Behavioural processes*, **157**, 402-407. https://doi.org/10.1016/j.beproc.2018.07.009

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

Chen, G., Pine, D. S., Brotman, M. A., Smith, A. R., Cox, R. W., Taylor, P. A., & Haller, S. P. (2021). Hyperbolic trade-off: the importance of balancing trial and subject sample sizes in neuroimaging. *NeuroImage*, 118786. https://doi.org/10.1016/j.neuroimage.2021.118786

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

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

Cohen, J. (2016). A power primer. In A. E. Kazdin (Ed.), Methodological issues and strategies in clinical research (pp. 279–284). American Psychological Association. https://doi.org/10.1037/14805-018

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

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

Couvillon, M. J., Jandt, J. M., Duong, N. H. I., & Dornhaus, A. (2010). Ontogeny of worker body size distribution in bumble bee (*Bombus impatiens*) colonies. *Ecological Entomology*, **35**, 424-435. https://doi.org/10.1111/j.1365-2311.2010.01198.x

Crawley, M. J. (2007). The R Book. John Wiley & Sons. *Chichester, UK*.

DeHeer, C. J., & Vargo, E. L. (2006). An indirect test of inbreeding depression in the termites *Reticulitermes flavipes* and *Reticulitermes virginicus*. *Behavioral Ecology and Sociobiology*, **59**, 753-761. https://doi.org/10.1007/s00265-005-0105-9

Engemann, K., & Wall, H. J. (2009). The effects of recessions across demographic groups. *Available at SSRN 1490041*. http://dx.doi.org/10.2139/ssrn.1490041

Frank, E. T., Wehrhahn, M., & Linsenmair, K. E. (2018). Wound treatment and selective help in a termite-hunting ant. *Proceedings of the Royal Society B: Biological Sciences*, **285**, 20172457. https://doi.org/10.1098/rspb.2017.2457

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

Grilli, L., & Rampichini, C. (2015). Specification of random effects in multilevel models: a review. *Quality & Quantity*, **49**, 967-976. https://doi.org/10.1007/s11135-014-0060-5

Hector, A., Hautier, Y., Saner, P., Wacker, L., Bagchi, R., Joshi, J., Scherer-Lorenzen, M., Spehn, E.M., Bazeley-White, E., Weilenmann, M. and Caldeira, M.C (2010). General stabilizing effects of plant diversity on grassland productivity through population asynchrony and overyielding. Ecology, 91(8), 2213-2220. https://doi.org/10.1890/09-1162.1

Hölldobler, B., & Wilson, E. O. (2009). *The superorganism: the beauty, elegance, and strangeness of insect societies*. WW Norton & Company.

Jandt, J. M., Bengston, S., Pinter‐Wollman, N., Pruitt, J. N., Raine, N. E., Dornhaus, A., & Sih, A. (2014). Behavioural syndromes and social insects: personality at multiple levels. *Biological Reviews*, **89**, 48-67. https://doi.org/10.1111/brv.12042

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

Johnson, P.C.D., Barry, S.J.E., Ferguson, H.M. & Müller, P. (2015) Power analysis for generalized linear mixed models in ecology and evolution. *Methods in Ecology and Evolution*, **6**, 133– 142. https://doi.org/10.1111/2041-210X.12306

Kain, M. P., Bolker, B. M., & McCoy, M. W. (2015). A practical guide and power analysis for GLMMs: detecting among treatment variation in random effects. *PeerJ*, **3**, e1226. https://doi.org/10.7717/peerj.1226

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

London, K. B., & Jeanne, R. L. (2003). Effects of colony size and stage of development on defense response by the swarm-founding wasp *Polybia occidentalis*. *Behavioral Ecology and Sociobiology*, **54**, 539-546. https://doi.org/10.1007/s00265-003-0662-8

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

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

Porter, E. E., & Hawkins, B. A. (2001). Latitudinal gradients in colony size for social insects: termites and ants show different patterns. *The American Naturalist*, **157**, 97-106. https://doi.org/10.1086/317006

Proust, C., & Jacqmin-Gadda, H. (2005). Estimation of linear mixed models with a mixture of distribution for the random effects. *Computer methods and programs in biomedicine*, **78**, 165-173. https://doi.org/10.1016/j.cmpb.2004.12.004

Masciocchi, M., Pereira, A. J., Lantschner, M. V., & Corley, J. C. (2013). Of volcanoes and insects: the impact of the Puyehue–Cordon Caulle ash fall on populations of invasive social wasps, Vespula spp. *Ecological research*, **28**, 199-205. https://doi.org/10.1007/s11284-012-1004-y

Messier, J., McGill, B. J., & Lechowicz, M. J. (2010). How do traits vary across ecological scales? A case for trait‐based ecology. *Ecology letters*, **13**, 838-848. https://doi.org/10.1111/j.1461-0248.2010.01476.x

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

Montgomery, D. C. (2013). *Statistical quality control*. Wiley Global Education.

Montgomery, D. C. (2020). *Design and analysis of experiments*. John wiley & sons.

Parr, L. A., Waller, B. M., & Vick, S. J. (2007). New developments in understanding emotional facial signals in chimpanzees. *Current Directions in Psychological Science*, **16**, 117-122. https://doi.org/10.1111/j.1467-8721.2007.00487

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL [http://www.R-project.org/](https://www.r-project.org/)

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

Schielzeth, Holger, Niels J. Dingemanse, Shinichi Nakagawa, David F. Westneat, Hassen Allegue, Céline Teplitsky, Denis Réale, Ned A. Dochtermann, László Zsolt Garamszegi, and Yimen G. Araya‐Ajoy. (2020). Robustness of linear mixed‐effects models to violations of distributional assumptions. *Methods in Ecology and Evolution*, **11**, 1141-1152. https://doi.org/10.1111/2041-210X.13434

Schulz, D. J., Sullivan, J. P., & Robinson, G. E. (2002). Juvenile hormone and octopamine in the regulation of division of labor in honey bee colonies. *Hormones and behavior*, **42**, 222-231. https://doi.org/10.1006/hbeh.2002.1806

Tay, W. T., & Crozier, R. H. (2001). Mating behaviour of *Rhytidoponera* sp. 12 ants inferred from microsatellite analysis. *Molecular Ecology*, **10**, 167-173. https://doi.org/10.1046/j.1365-294X.2001.01167.x

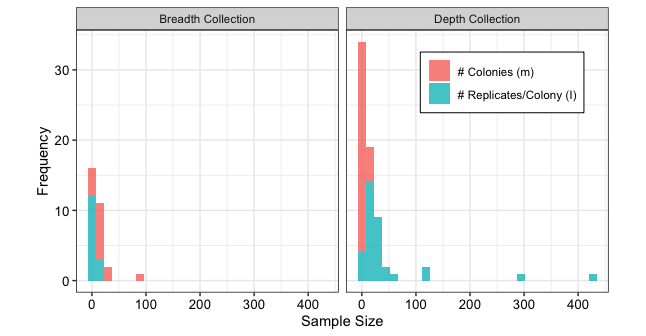
Udino, E., Perez, M., Carere, C., & d’Ettorre, P. (2017). Active explorers show low learning performance in a social insect. *Current zoology*, **63**, 555-560. https://doi.org/10.1093/cz/zow101

Vargha, A., & Delaney, H. D. (1998). The Kruskal–Wallis test and stochastic homogeneity. *Journal of Educational and behavioral Statistics*, **23**, 170-192. https://doi.org/10.3102/10769986023002170

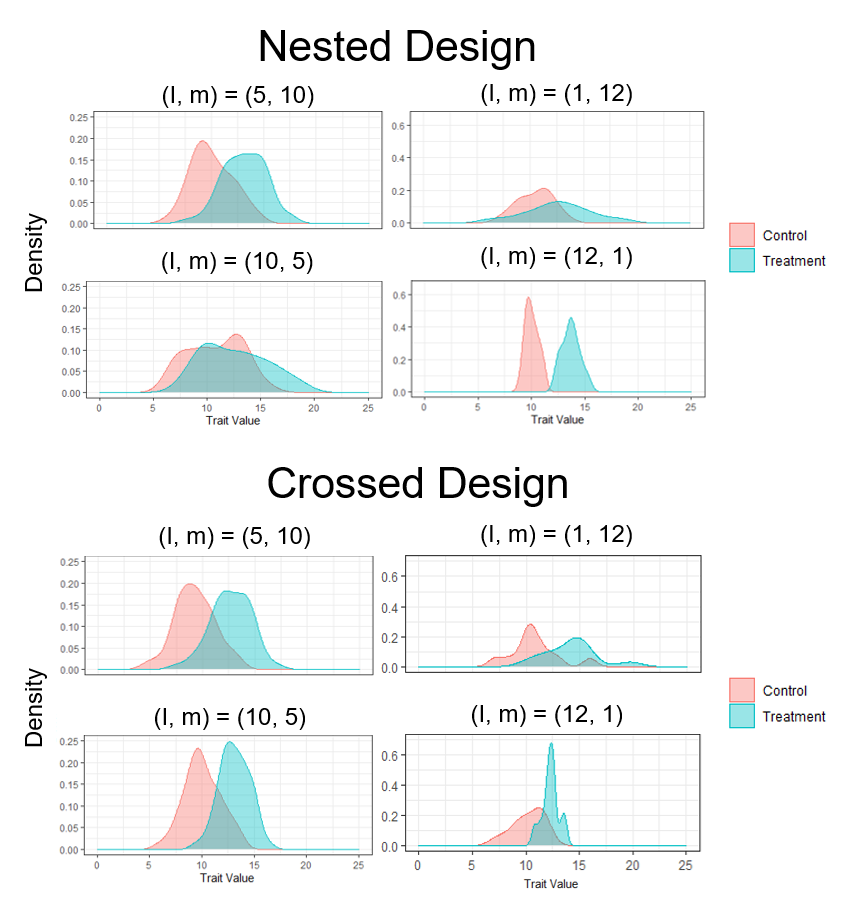
Wang, Y. A., Sparks, J., Gonzales, J. E., Hess, Y. D., & Ledgerwood, A. (2017). Using independent covariates in experimental designs: Quantifying the trade-off between power boost and Type I error inflation. *Journal of Experimental Social Psychology*, **72**, 118-124. https://doi.org/10.1016/j.jesp.2017.04.011

Wright, C. M., Lichtenstein, J. L., Doering, G. N., Pretorius, J., Meunier, J., & Pruitt, J. N. (2019). Collective personalities: present knowledge and new frontiers. *Behavioral Ecology and Sociobiology*, **73**, 1-23. https://doi.org/10.1007/s00265-019-2639-2

**Supplementary Information**



**Figure S1.** Distributions of sample sizes across collection types and types of samples. Color gives the collection type, the x-axis gives the sample size, and the y-axis is count. Sample sizes are given as the number of samples per treatment. The x-axis was limited for visibility, as there was one case where the number of within-colony replicates was greater than 3,000.



**Figure S2.** Example simulations for each type of design with different collection strategies. In these simulations, all free parameters are equal to 3. Each plot shows the kernel density of values present in the control group (pink) or the treatment group (blue). Nested designs with depth collection tend to exaggerate the mean shift for small values of n, but this effect disappears at larger values of n or when breadth collection is used instead. Crossed designs tend to recreate control and treatment distributions faithfully regardless of sample size.

**S 1.1 Normal versus t distributions**

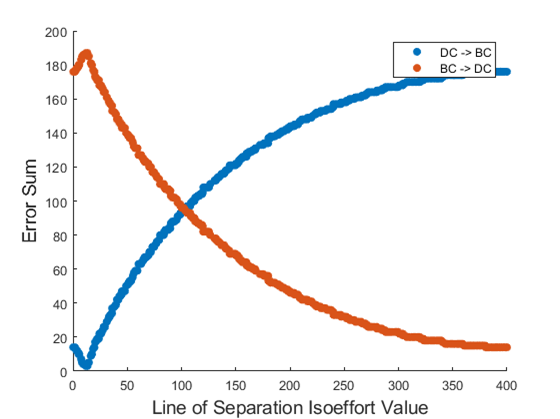
The sum of two normally distributed random variables is also normal, so for large sample sizes one may use a z-test to differentiate between the two treatments (if one ignores the effect of colonies). However, for small sample sizes (typically below 30), a t-distribution is used as the reference distribution instead of the normal distribution (specifically for t-tests, Montgomery, 2020). This divergence is likely key to several results of this study, so we measure the relative ability of each type of distribution to fit sampled data.

After a simulation (Text 2.2), the data from both the control and treatment groups are pooled and compared to both z and t distributions with a one sample K-S test. The mean for both reference distributions is simply the mean of the dataset and the variance is set to the variance of the dataset. The degrees of freedom for the t-distribution is equal to the sample size - 1. If the test statistic for the K-S test is smaller for one distribution over the other and p < 0.05, then that distribution has a better fit for the data. If the p-value for both tests was greater than 0.05, then the two were considered a draw. We counted the number of the cases in which t distribution provided a better fit over the normal distribution while ignoring draws. We refer to this index as t-count.

**S 1.2 Minimal sample size per treatment for nested designs**

The relative performance of breadth and depth collection strategies changed across the sampling space in nested designs. For instance, depth collection strategies had higher balanced accuracies at smaller sample sizes (low values of n) while breadth collection strategies performed better at larger sample sizes. This clustering effect is distinct enough that we can find the isoeffort curve above which breach collection always performs better than depth collection, and vice versa below the curve. This isoeffort contour corresponds to a fixed value of n, so W = 1 while E = n.

To find the line of best separation between these two clusters, we performed two ‘sweeps’ through the sampling space. First, we estimated balanced accuracy for all possible sampling strategies within the given sampling space (Text 2.2). Then in the first ‘sweep’ we draw all possible isoeffort contours for n ∈ {1, 2, …, 400}, and then classified all points below that line (except those that correspond to balanced collection strategy) as “breadth collection balanced accuracy > depth collection balanced accuracy” and all points above that line as “breadth collection balanced accuracy < depth collection balanced accuracy.” Comparisons are made across conjugate pairs, so the balanced accuracy of the breadth collection strategy (2, 4) is compared to the breadth collection strategy (4, 2). Error in this context is the number of misclassified points across the sampling space for a particular isoeffort curve. In the second sweep, we undergo the same process, but reverse the classification scheme. The line of best separation has the minimum error for both these sweeps (Fig. S3).



**Figure S3.** The line of separation determines the isoeffort curve used to separate the sampling space. DC = Depth Collection and BC = Breadth Collection. One set of thresholds sets all the points below the isoeffort curve to DC > BC and above it to BC > DC (blue points, DC -> BC), while the other type of threshold does the opposite (red points, BC -> DC). The error sum gives the count of all points that are misclassified on both sides of the line. Here, the global minimum for balanced accuracy across all parameters occurs along the DC -> BC line at n = 14 (N = 28). Balanced accuracy here is pooled across all parameter combinations.

First, we found the line of best separation for when balanced accuracy is estimated across all parameter combinations. Here, the line of best separation is n = 14. However, we also found the lines of best separation for each individual parameter combination. In many cases, clustering was not as clean as it is for the pooled case. Some points above or below the line could have a mixture of points where breadth collection outperformed depth collection and vice versa.

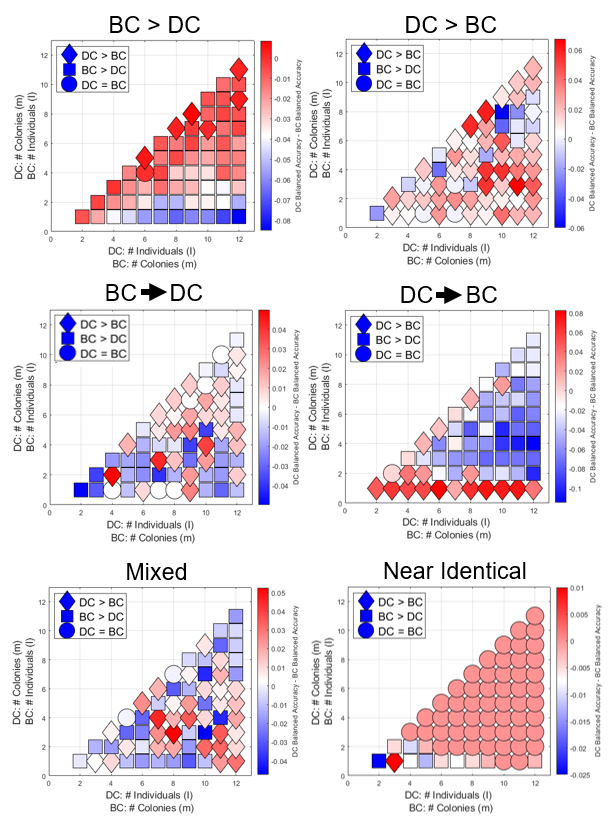
To measure the degree of agreement above and below the line, we calculate a value called homogeny. Homogeny is the count of the dominant strategy above or below the line divided by the total number of potential strategies above or below the line. Homogeny is calculated separately for the regions above and below the line. The final homogeny for the whole sampling space is the weighted sum of these two homogenies. Weights are determined by the number of potential strategies above or below the line.

**S 1.3 Sampling space categories**

The sampling space for any particular combination of the parameters Δ, σw, and σa may or may not be divisible as it is in the aggregate case for nested designs (Fig. 2A). In some situations, one strategy may be dominant over another regardless of the sample size. In other situations, neither strategy is dominant, so their successes are mixed across the sampling space. Situations where depth collection has a higher balanced accuracy at low sample sizes, and breadth collection dominates at high sample sizes are labeled as ‘DC -> BC’, the reverse is ‘BC -> DC.’ Situations where depth collection dominates over the full sample spectrum is ‘DC > BC,’ and the reverse is ‘BC > DC.’ Situations where results are mixed are simply titled ‘mixed’ and situations where both depth and breadth collection yield a balanced accuracy of 1 at most sample sizes is called “near identical.” There are also occasions where small sample sizes are best described as ‘DC -> BC’ or ‘BC -> DC,’ but at large sample sizes both depth and breadth collection yield a balanced accuracy of 1, so these situations are labeled ‘DC -> BC -> Identical’ or ‘BC -> DC -> Identical.’

There are 1,000 combinations of parameters, so we need an automated sorting system to distinguish between these situations. We first classified all situations as being either DC -> BC or BC -> DC using the scheme presented in S 1.2. We then measure homogeny (see S1.3), the proportion of cases where BC has a higher balanced accuracy than DC, and the proportion of cases where the reverse is true. If homogeny is below 0.8 and the proportion of BC cases is higher than 0.75, then the situation was classified as BC > DC. Alternatively, if the proportion of BC cases is higher than 0.85, it would also be classified as BC > DC. The same applies for DC. If homogeny is below 0.8 and the proportion of DC cases is higher than 0.75, then the situation was classified as DC > BC. If homogeny is below 0.6, then it was classified as mixed. If in 50% or more of cases the difference between breadth and depth collection strategies was 0 (because both balanced accuracies are 1) , then it was classified as nearly identical. Finally, if a sampling space has already been classified as ‘BC -> DC’ or as ‘DC -> BC’ and 10% or more of cases the difference between breadth and depth collection strategies was 0 (but less than 50%), then it was classified as ‘BC -> DC -> Identical’ or ‘DC -> BC -> Identical.’

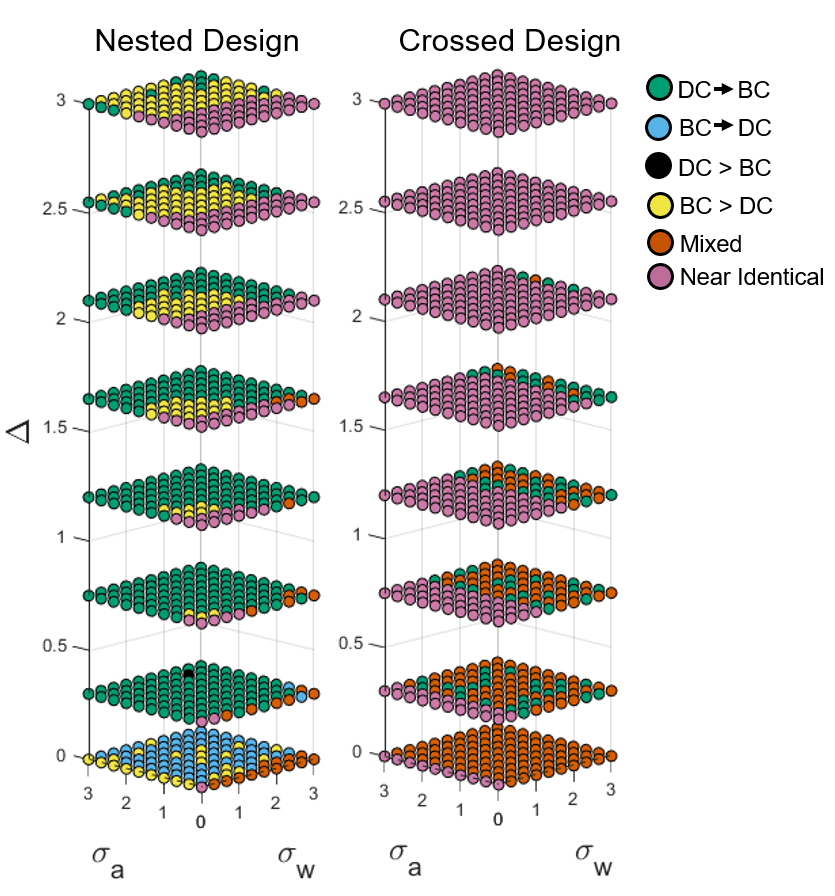
These threshold values were chosen somewhat arbitrarily. They were manually adjusted until the classification scheme matched what the observer would have adjusted the scheme. The final scheme was tested with 40 randomly selected parameter values, and of these 40, 34 were classified correctly. The other 6 were ambiguous but not strictly incorrect (see Fig. S4 for examples of each category).



**Figure S4.** Each plot gives an example sampling space for each of the categories used in Figure 4. How these categories were selected is discussed in the section “Sampling space categories.” Here we limit the sampling space such that *I, m* < 13, as it is easier to visualize this smaller space than the larger *I, m* < 21 space.

**S 1.4 Sampling space for various sampling techniques: effect of free parameters**

The free parameters Δ, 𝛔a, and 𝛔w can have an effect on the relative advantage of depth and breadth collection across different sample sizes. In crossed designs, sample space remains symmetric across parameter combinations (Fig. S5). On the other hand, in nested designs, the DC -> BC pattern was observed in Fig. 2A could be variable in some parameter combinations (Fig. S8). When there was no effect of treatment (Δ = 0), the performance of a sampling strategy is measured by reducing false positive rate. In this case, the pattern was opposite to Fig. 2A: BC achieved higher balanced accuracy than DC at small sample sizes. This happens because the distance between the control and treatment distributions is smaller for BC over DC. This means that there are fewer false positives for BC than there are for DC. However, for larger values of n, the difference in the false positive rate between BC and DC disappears so the category for the sampling space gets encoded as BC -> DC. At small, nonzero values of Δ, DC starts to perform better, as it can amplify the signal and detect a true shift, especially at low sample sizes.

****

**Figure S5.** Plot gives the 3-D parameters space of nested designs where the line of best fit thresholds tend to cluster for both the nested design (A) and the full factorial design (B). Each axis is a parameter, and color describes how that sampling space is organized. Situations where depth collection is preferable at low sample sizes while breadth collection is better at larger sample sizes are labeled as ‘DC -> BC’ while the opposite is ‘BC -> DC.’ Situations where depth collection dominates over the full sample spectrum is ‘DC > BC,’ and the reverse is ‘BC > DC.’ Situations where neither depth nor breadth collection dominate at any sampling size are called ‘mixed’ and situations where the difference between breadth and depth collection is 0 is called ‘near identical.’ There are also occasions where small sample sizes are best described as ‘DC -> BC’ or ‘BC -> DC,’ but at large sample sizes both depth and breadth collection yield a balanced accuracy of 1, so these situations are labeled ‘DC -> BC -> Identical’ or ‘BC -> DC -> Identical.’

**S 1.5 Relative performance of sampling strategies using power and balanced accuracy**

We compared the balanced accuracies of different experimental designs (nested and crossed) or different sampling strategies (depth, balanced, and breadth collection), using Kruskal–Wallis tests. We used Dunn’s post-hoc for pairwise comparisons in case of significant differences (Vargha & Delaney, 1998). The performance was compared by eight measurements, power, balanced accuracy, false-positive rate, false-negative rate, square error of estimates of parameters (Δ, σw, and σa), and the number of better fits to t-distribution over normal distribution (Fig. 3D). All these variables were not normally distributed (Shapiro–Wilk test, p < 0.001 in all cases) and generally heteroscedastic (Bartlett test p < 0.01 in most cases). We performed these statistical tests on 5,000 randomly sampled data points to limit computational overhead.

Crossed designs generally had a higher balanced accuracy than nested design (pool all parameters: one-sided Mann-Whitney U test W = 7,543,964, p < 0.001). In crossed designs, the balanced accuracy did not differ among three sampling strategies, breadth, depth, and balanced collection (Kruskal–Wallis test, χ2 = 0.129, p > 0.05, Fig. 2B). When l or m = 1, breadth collections were slightly preferred over depth collections (Fig. 2B). However, in general, the balanced accuracy did not change across sampling strategies for any given value of n (Fig. 2B). Consistent with this, there was no difference in power or true negative rates among sampling strategies (Kruskal–Wallis tests, p > 0.05, Fig. 2DE).

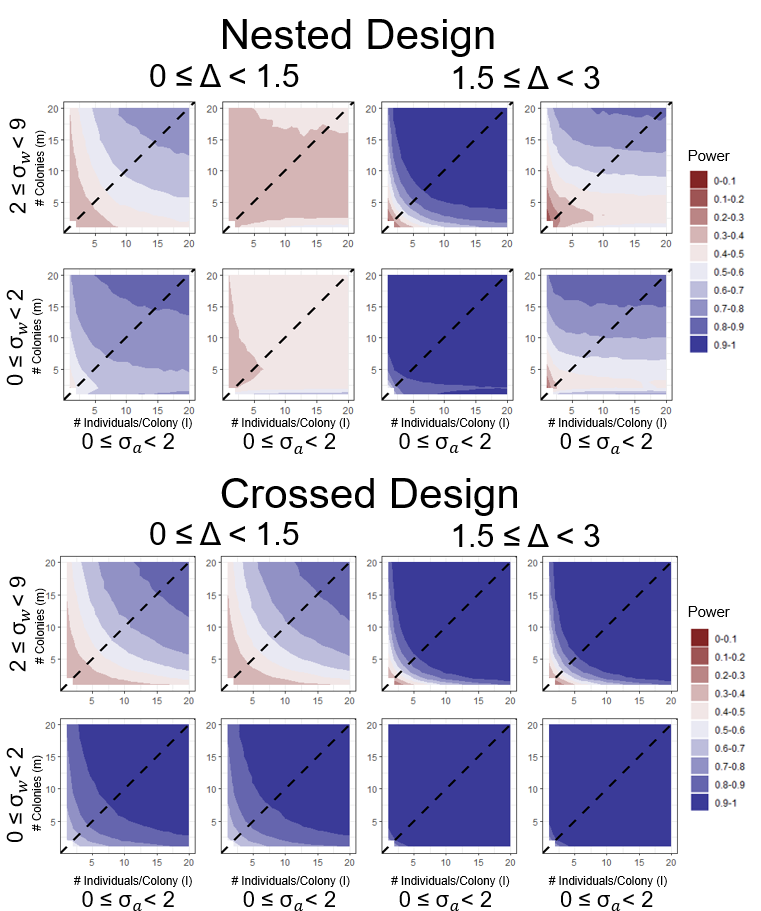
We also evaluated how power changes with design and sampling strategy. Crossed designs generally outperformed nested design (pooled across all parameters: one-sided Mann-Whitney U test W = 4,069,123, p < 0.001). In crossed design, balanced accuracy did not differ among three sampling strategies, breadth, depth, and balanced collection (Kruskal–Wallis test, χ22 = 0.51627, p > 0.05). When l or m = 1, breadth collection is preferred over depth collection. In general, power only increases with n, whether samples were allocated to increase within-colony replicates or among-colony replicates (Fig. 2BD). Consistent with this, there was no difference in estimated parameters or false positive (negative) rate among sampling strategies (Kruskal–Wallis tests, p > 0.05).

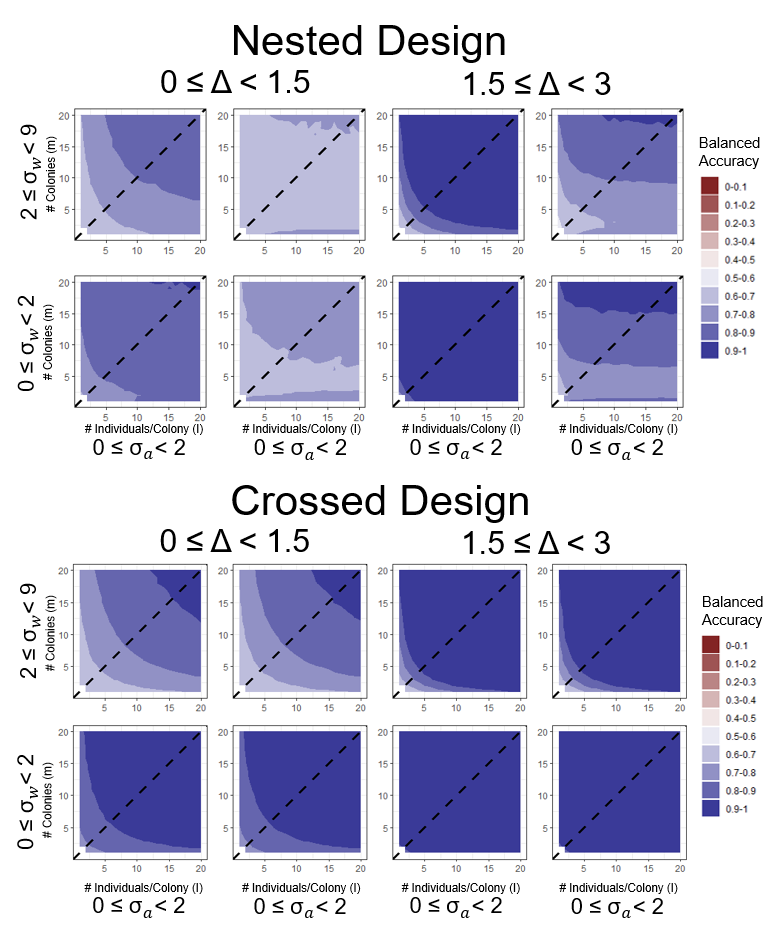
On the other hand, power depended on sampling strategies in nested designs. Breadth collection generally achieved higher power than depth collection, where the power increased with more among-colony replicates, rather than more within-colony replicates (Kruskal–Wallis test, χ2 = 18.8934, post-hoc Dunn test with Holm correction p < 0.001 comparing depth and breadth, p comparing breadth and depth to balanced > 0.05). However, at small sample sizes (n < 20), sampling more within-colony replicates than the number of colonies leading to higher power (Fig. 3AC, Text S1.3, Fig. S6). This occurred because the false negative probability for depth collection is smaller at lower sample sample sizes than it is for breadth collection, but this relationship flips for n ≥ 20.

**S 1.6 Balanced accuracy across sampling space and all sampling techniques**

Parameter values have a strong effect on balanced accuracy as well as power. While different response surfaces will arise in the sampling space from different combinations of free parameters (see Text S1.5), pooling their effects within various ranges of the parameters will not qualitatively affect the patterns described in the fully pooled dataset (Fig. 2, Fig. S6, Fig. S7). If a reader has a general idea of how the parameters relate to each other (for instance, they may know if inter colony variation is higher than intra colony variation), then they can refer to figure S6 for an estimate of power across different parameter combinations or figure S7 to estimate balanced accuracy.

To make comparisons between crossed designs, nested designs, depth collection, breadth collection, and balanced collection for different parameter combinations, we pooled the data into 8 parameter regions (where combinations of parameter values are either low or high, 23 = 8), and then performed a Kruskal-Wallis test within each region. All tests were significant (all p values < 0.001), so we also performed a post-hoc Dunn test within each region to make pairwise comparisons. In all cases, crossed designs had a higher balanced accuracy and power than nested designs regardless of the parameter space. Within crossed designs, there tended not to be any differences in balanced accuracy or power between depth, breadth, and balanced collection. Within nested designs, balanced and breadth collection outperformed depth collection, however they tended not to be different from one another.

**Figure S6.** Power contour plots for each type of sampling across the parameter space. The top set of plots were generated from nested designs, while the bottom set was generated from crossed designs. Plots represent the sampling space, and color gives balanced accuracy. Each parameter was divided evenly into high and low levels, giving 8 combinations of parameters, and the data presented is the pooled average accuracy within a particular parameter space.



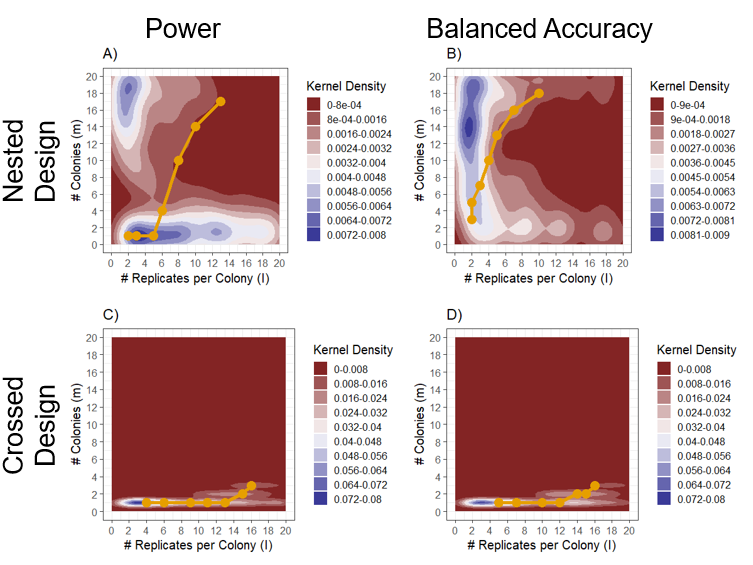
**Figure S7.** Balanced accuracy contour plots for each type of sampling across the parameter space. The top set of plots were generated from nested designs, while the bottom set was generated from crossed designs. Plots represent the sampling space, and color gives balanced accuracy. Each parameter was divided evenly into high and low levels, giving 8 combinations of parameters, and the data presented is the pooled average accuracy within a particular parameter space.

**S 1.7 Distribution of optimal sampling strategies across free parameters**

The free parameters Δ, 𝛔a, and 𝛔w also have an effect on the optimal sampling strategy. As of yet, there are no scientific standards on the amount of effort one should put into a study, but there are standard power and significance levels used throughout the biological sciences. We therefore find the distribution of optimal strategies by minimizing effort given lower limits on power (0.8) or balanced accuracy (0.9) for both nested and crossed designs rather than using effort as an upper limit. We then calculate balanced accuracy or power for all (*I*, *m*) combinations as well as effort (W = 2.94) for each parameter combination (except where 𝛔a or 𝛔w = 0, as these are biologically infeasible). We used kernel density estimation to approximate the probability density function of optimal strategies (Fig. S8).

Patterns in this distribution mirror patterns discussed in the example for Text 3.3. Power tends to favor depth collection for nested designs, but balanced accuracy correctly supports breadth collection. Conversely, power and balanced accuracy give nearly the same answers for crossed designs.

We also show different strategies that a researcher could use if the underlying parameter space of their model system is unknown. Gold points show various percentiles for I and m across optimal sampling strategies. In other words, these are roughly the percent chances a researcher would get power > 0.8 or balanced accuracy > 0.9 if they do not know the underlying parameter space of their experiment. If a researcher wants to be more conservative and assume that they have a small signal-to-noise ratio, they can select higher percentile strategies which have a stronger probability of minimizing error. If they do not have the resources to make this assumption, then they can choose lower percentiles.

****

**Figure S8.** Each plot shows the distribution of optimal strategies for nested (A, B) and crossed (C, D) designs according to either power (A, C) or balanced accuracy (B, D). Color gives the kernel density of points, where blue shows regions which are more concentrated and red shows sparse regions. Gold points show percentiles for I against percentiles for m. From left to right, these points show 20%, 30%, …, 80% percentiles. In other words, these are the percent chances a researcher would get power > 0.8 or balanced accuracy > 0.9 if they do not know the underlying parameter space of their experiment.

**ORCID**

C.L.: 0000-0001-7238-6801

T.P.P.: 0000-0002-7073-6932

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

**Acknowledgements**

We thank Kaitlin Baudier, Jennifer Fewell, and members of the Fewell and SEADS labs for helpful discussion.

NM is supported by JSPS Overseas Research Fellowships and JSPS Research Fellowships for Young Scientists, SPD and CPD (grant number: 20J00660).

CL is supported by NSF GRFP.

**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.

**How does this fit the journal’s scope?**

The statistical methods presented in this manuscript will assist social insect scientists in their sampling methodologies. Further, as it considers sampling across a random effect, the work presented here represents a general framework that scientists from other disciplines could use as well.

**Conflict of Interest**

The authors declare no conflict of interest.

**We intend to archive the data on the Dryad database (https://datadryad.org/stash).**

<https://besjournals.onlinelibrary.wiley.com/hub/journal/2041210X/author-guidelines>