Journal of the American Association for Laboratory Animal Science
Copyright 2020
by the American Association for Laboratory Animal Science

Vol 59, No 1
January 2020
Pages 9–16

Power to the People: Power, Negative Results and Sample Size

Brianna N Gaskill,[1,*] and Joseph P Garner,[2]

The practical application of statistical power is becoming an increasingly important part of experimental design, data analysis, and reporting. Power is essential to estimating sample size as part of planning studies and obtaining ethical approval for them. Furthermore, power is essential for publishing and interpreting negative results. In this manuscript, we review what power is, how it can be calculated, and reporting recommendations if a null result is found. Power can be thought of as reflecting the signal to noise ratio of an experiment. The conventional wisdom that statistical power is driven by sample size (which increases the signal in the data), while true, is a misleading oversimplification. Relatively little discussion covers the use of experimental designs which control and reduce noise. Even small improvements in experimental design can achieve high power at much lower sample sizes than (for instance) a simple *t* test. Failure to report experimental design or the proposed statistical test on animal care and use protocols creates a dilemma for IACUCs, because it is unknown whether sample size has been correctly calculated. Traditional power calculations, which are primarily provided for animal number justifications, are only available for simple, yet low powered, experimental designs, such as paired *t* tests. Thus, in most controlled experimental studies, the only analyses for which power can be calculated are those that inheriently have low statistical power; these analyses should not be used because they require more animals than necessary. We provide suggestions for more powerful experimental designs (such as randomized block and factorial designs) that increase power, and we describe methods to easily calculate sample size for these designs that are suitable for IACUC number justifications. Finally we also provide recommendations for reporting negative results, so that readers and reviewers can determine whether an experiment had sufficient power. The use of more sophisticated designs in animal experiments will inevitably improve power, reproducibility, and reduce animal use.

DOI: 10.30802/AALAS-JAALAS-19-000042

Deciphering negative results

In September of 2017, JAALAS published an editorial on its openness to publishing negative results.[47] The editors of the journal believe that this information is essential to the field of laboratory animal science for making evidence-based-decisions, even when comparisons made in a study do not show statistically significant differences. However, reviewers and readers should consider 2 possibilities when interpreting negative results: either 1) the treatment truly had no effect on experimental outcomes; or 2) the treatment has an effect, but the effect was not detected due to issues with the experimental design, issues with the experimental analysis, or random chance. Teasing out which of these possibilities is true requires us to consider the power of the original experiment.

We concur with the original editorial's intent to support evidence-based decision making and with the importance of reporting negative results, which is particularly important for addressing the reproducibility and translatability crises.[2,3,11,12,18] This is true for both of the negative result scenarios presented above. Imagine a situation in which a hypothesis is exciting, which results in many research groups working in parallel, each trying to scoop the other, but the hypothesis is fundamentally false. Truly negative results may never be reported, but sooner or later an experiment will be significant simply by chance. This false discovery ends up being the first result published (even though the previously completed negative experiments are unpublished). These false discoveries are most often the product of artifacts from experimental design, housing, test conditions, and other unrecognized factors.[12,18,39,40] However, identifying this discovery as false may not occur for years because it is highly replicable in its home lab.[2,3,48] This false but enthralling idea can snowball because other groups are convinced by the first lab's string of positive results. These groups continue on with ever-larger studies, all chasing the same result, only publishing statistically significant (but scientifically incorrect) results, and wasting large amounts of time, money, and animals (for a powerful example in human research, see[43]).

Conversely, imagine the second scenario, in which a hypothesis is correct but is rather unexciting scientifically. In this case, initial studies may be small and underpowered. Negative results will be unpublished until eventually one experiment correctly finds a treatment effect and is published. In underpowered experiments, an effect has to be larger than the true effect size to be significant.[5] As a result, in many fields, as a finding is replicated with multiple larger, more highly-powered experiments, the consensus effect size decreases and approaches its true value.[5] Both scenarios hinder or slow the progress of science, and both could be avoided if the original experiments were properly powered and negative results are reported.

We use inferential statistics as a tool to draw conclusions about questions regarding natural phenomena. In its simplest form, statistics is used to compare an estimate to something that is known. For instance, when we calculate a *P* value, we

Received: 25 Mar 2019. Revision requested: 02 May 2019. Accepted: 24 Sep 2019.
[1]Animal Sciences, Purdue University, West Lafayette, Indiana; [2]Comparative Medicine, Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford California
[*]Corresponding author. Email: bgaskill@purdue.edu

assume that the null hypothesis (H_0) is true (that is this is what is known) and we estimate the chance that an effect as large as the one observed would occur. Thus when we calculate *P* we are essentially asking the question '*given that the null hypothesis is true, what proportion of experiments would yield a result this unusual?*'. Most importantly, *P* values are meaningless without a null hypothesis, (see Figure 1). The power of a specific statistical test is the probability that it will accurately reject the null hypothesis (H_0) in favor of the alternative (H_A) when the null hypothesis is actually false.[11,29] Thus power asks the question '*given that the alternative hypothesis is true, what is the chance of seeing a significant result?*' (Figure 1.)

Power as an engineering standard

Traditional calculations of power and sample size, and their pitfalls. Power is traditionally calculated in the planning stages of a study to determine the sample size (for example, the number of animals) that will be needed to achieve an acceptable level of power, and such justifications of sample size are required for IACUC protocol approval. (Note that sample size should never be justified by habit—'*We have previously used 10 animals*'—this is not considered a valid calculation of sample size).[15] These calculations are referred to as 'prospective power,' 'prior power,' or 'a priori power.'[45] A central point of this overview is that while this is fine in theory, in reality, insisting or recommending traditional power calculations as the method for sample size justification is neither good science nor good animal welfare. This is because there is a conflict between the experimental designs that maximize power and the analyses for which power is calculable. In other words, while it may be possible to calculate power for analyses such as *t* tests or one-way ANOVAs, these analyses require orders of magnitude more animals than more complex approaches (like randomized block designs) for which no formal power analysis is available.[10,11] While some user-friendly tests[24] exist to help researchers determine power; the facts remain that power primarily depends on experimental design, and that the ultimate test being used to analyze experimental results is not usually discussed. Furthermore, the entire notion of power calculation is fraught with difficulty, all of which creates frustration and uncertainty for researchers and IACUC alike. In this section we explain traditional power calculation theory and then move on to explain why this approach is so problematic. We will then offer practical and simple solutions to this dilemma.

For tests where power is calculable, power equations describe the relationship between 4 variables: 1) significance criterion or critical α (acceptable Type I error rate; generally $\alpha < 0.05$); 2) the sample size, N; 3) the population effect size; and 4) power.[17] All of these variables are set or estimated '*before*' the experiment is performed. Although we refer to this as an 'a priori power calculation', in reality, any variable can be calculated if the others are known. The significance criterion is generally accepted at critical $\alpha = 0.05$. Power can be figured for a given effect size and sample size, but in most situations, power is a set at a minimally acceptable level (typically 80 or 90%—that is assuming H_A is true, we want to ensure an 80 or 90% chance that an experiment will reveal this result). From there, we can calculate the Least Significant Number (that is the minimum sample size required for a given effect size; LSN), or the Least Significant Value (that is the effect size that would just achieve significance for a given sample size). When planning experiments, we are typically interested in determining the LSN, and so the final piece of information required is the effect size of biologic interest. This can be easier to estimate in some types of studies, such as clinical trials because response scores often have mild, moderate, and severe classifications that may provide a numerical value of clinical interest. Alternatively, we can use previous literature to determine effect size. However, it is not general practice in laboratory animal science to report effect sizes, as is done in human research, and prior observed effect sizes are often misleading.[5]

Effect size is typically expressed as some standardized measure of how much the outcome variable is affected by predictor variables, relative to error. This value conveys the magnitude of difference found between groups in an easily understood scale.[34,44] While effect size communicates how much a treatment will change a measure, this is different from statistical significance (such as calculating a *P* value from an *F* or *t* test), which estimates whether the difference found is believably due to the treatment.[44] Further, the degree of significance ($P = 0.03$ compared with $P = 0.0001$) does not necessarily mean that the effect size is large.[44] Essentially, statistical significance tells you whether a difference is detectable, while effect size tells you how big of a difference it is. Thus effect size is particularly important because it allows you to assess whether a statistically significant difference is biologically meaningful.

When calculating effect size, such as Cohen *d*, you would determine the mean difference between the treatment and control group, divided by their pooled standard deviation.[44] Cohen proposed that *d* values of 0.2, 0.5, and 0.8 are considered small, medium, and large effect sizes.[22] These values are fairly arbitrary and do not take into account other important variables such as accuracy of the measurements or variability within the study population.[34,44] Both *F* and *t* values are also effect sizes (though this is poorly appreciated): *F* is the variance due to the treatment divided by variance due to error; whereas *t* is the mean difference divided by its standard error ($F = t^2$ for effects with 1 degree of freedom; DF).

Most readers are probably familiar with R^2, or the proportion of the total variation in the response variable that can be accounted for by the analysis.[6,38] R^2 is traditionally used as a measure of how well the statistical model fits the data (for example if $R^2 = 0.68$, then 68% of the variation in the data can be explained by treatment and other predictors in the model, and 32% is unexplained noise). Once multiple predictors are in an analysis (for example, imagine an analysis predicting mouse body weight, given both age and genotype), then each predictor is tested given the other predictors in the model. The corresponding *F* and *P* values test the unique explanatory power that each predictor adds to the model. The corresponding measures of effect size are now referred to as "partial". For continuous predictors, like age in the example above, we calculate partial R^2 (which measures the proportion of variation in the data uniquely explained by the predictor). When meaningful, this can be converted to a partial correlation coefficient.[6] For categorical predictors (like genotype in the example above), we calculate partial eta squared (η_p^2) exactly as we would partial R^2. Thus both partial R^2 and partial eta squared measure the proportion of variation in the data that is uniquely explained by the predictor. As with many things in statistics, this confusing and arbitrary terminology is a hangover from when different kinds of analyses were performed by hand. In more complex modern analyses, where we might want to test interactions,[35] and the effect size cannot be calculated as a simple difference or correlation coefficient, partial eta squared can always be calculated.

Furthermore, because partial eta squared has a consistent interpretation as the '*proportion of variation in the outcome variable explained*' across General Linear Models and their derivative tests (including *t* tests, ANOVA, regression, and correlation),

Term	Definition
Null hypothesis (H_0)	There is no meaningful difference between the populations being sampled.
Alternative hypothesis (H_A)	There is a meaningful difference between the populations being sampled that can be specified as a minimum numerical value.
Type I error	Detecting a positive 'significant' difference even though the null hypothesis is true.
Alpha (α)	Alpha is the probability of a Type I error *before* the experiment is performed. Typically, we choose a critical α that we are willing to tolerate *assuming* the null hypothesis is true: *"Given that the null hypothesis is true, what proportion of experiments will incorrectly (by chance) yield a significant result?"*
Significance (P)	P is the estimated chance of Type I error *after* the experiment is performed. P is compared against α to test significance, and thus also assumes that the null hypothesis is true: *"Given that the null hypothesis is true, what is the chance of seeing a result this weird?"*
Type II error	Detecting a negative result when the alternative hypothesis is true.
Beta (β)	Beta, like α, is the probability of a Type II error *before* the experiment is performed. In most cases, we talk about power (1-β), rather than β itself.
Power (1-β)	Power is the proportion of experiments that will produce a significant result assuming the Alternative hypothesis is true. Like α, power is figured or set *before* the experiment: *"Given that the alternative hypothesis is true, what is the chance of seeing a significant result?"*

Figure 1. Power-related terminology

and because partial eta squared can be converted into all other measures of effect size as long as the DF are known,[6,46] we and other authors[38] recommend that effect sizes always be reported as partial eta squared (a very user friendly tool for converting between effect sizes can be found at www.psychometrica.de/effect_size.html).[27] Reporting these effect sizes can help facilitate meta-analysis when interpreting or transcribing statistical tests during literature review, while also helping researchers conducting similar studies to easily calculate their needed sample size in a traditional power test. For partial eta squared (and thus also partial R^2), small, medium, and large effect sizes are considered 0.01, 0.06, and 0.14.[7,38] However, these categories should be treated with caution because they are essentially arbitrary and a result that is considered small in one field may be considered large in another.

Effect sizes can be estimated with pilot studies.[32] However, we and others[10,22] generally recommend against pilot studies, for both statistical and practical reasons. Pilot studies tend to be small, thus an unusually large effect is required to be significant, and thus pilot studies tend to overestimate the true effect size. Subsequent larger studies will tend to find smaller effects that approach the true effect size—a phenomenon called the 'winner's curse' or 'regression towards the mean'.[5] Furthermore, because the pilot typically overestimates the effect size, any power calculations using pilot observations will be incorrect and will advocate for too small a sample size in subsequent studies. In other words, pilot studies are unhelpful, because they actually increase the chance of a false negative; that is, they increase the chance of reproducibility failures. Conversely a well-designed and well-powered study can typically be performed for little additional effort.

The fallacy of post-hoc power

Imagine that an experiment is performed, and no significant result is observed. It is possible that the given negative result could be true (there is no effect), or the result may be a false negative (the effect exists, but it wasn't detected in this experiment). So how do we interpret this negative result? We have previously defined power as '*If the alternative hypothesis is true, what is the chance of seeing a significant result?*'. So, is it possible to tell these 2 scenarios apart using inferences from power? Some journals require that authors provide a power calculation as a condition for allowing negative results to be published, and as we explain below, some inference can be drawn from reporting a priori power calculations. This requirement becomes problematic, however, if the journal requests, or authors provide, an estimate of the power that was achieved in this single study. This power calculation is referred to as 'observed', 'achieved', 'retrospective', or 'post-hoc power'.[45] The use of retrospective power is a legitimate statistical assessment under very limited circumstances (as in the case of meta-analytical reviews of statistical power[5,45]), but never in the interpretation of a single negative result.[4,17,29,32]

Because power is the probability of an experiment correctly identifying H_A, power is only meaningful before an experiment is performed. Post-hoc power can be calculated by plugging the observed effect size, P values, and sample size into an a priori power formula, but the result is meaningless. If a result is significant, then post-hoc power will be high; if a result is not significant, then post-hoc power will be low.[17,29,45] This is neatly illustrated by the fact that if the observed P value = critical α (for example if we set critical $\alpha = 0.05$; and our experiment gives exactly $P = 0.05$), then post-hoc power = 50%.[5] This is true regardless of the critical α, so it is incorrect to assume that requiring more stringent critical α (for example $P < 0.01$) will increase post-hoc power. Thus a null result (regardless of critical α) inherently does not have sufficient post-hoc power. However, this does not mean that the null hypothesis was true, or that the experiment was actually underpowered. It simply means that the H_0 cannot be rejected.[14,36] The use of post-hoc power to (incorrectly) try to infer whether a nonsignificant result is a true-negative or a false-negative is referred to as the 'post-hoc power fallacy'.[17]

Reporting options if a nonsignificant result is obtained

The post-hoc power fallacy doesn't mean that negative results cannot be interpreted, just that post-hoc power is not helpful in doing so. A variety of options exist for meaningful interpretation of negative results. When a null result is reported, reviewers can consider whether a reasonable a priori sample size estimate was presented and whether the study met the target sample size.[29] Reporting a priori power, or sample size calculations, are recommended by the ARRIVE guidelines.[21] However, a recent publication tracking the implementation of these guidelines does not appear to have improved reporting.[28] Specific information regarding sample size, such as the total number of animals used and an a priori sample size calculation, were the worst reported sections of the guidelines in veterinary and animal welfare journals (19 out of 236 papers), which included publications from JAALAS and Comparative Medicine.[28] In a separate study conducted after Nature journals began requiring that a reporting checklist similar to the ARRIVE guidelines be submitted at publication, the reporting of sample size calculations improved by 56%.[31] Simply recommending or supporting the guidelines may not be enough to influence change in our field.

In line with other authors, our recommendation for reporting negative results is to include confidence intervals and observed effect sizes, along with test statistics (for example F values, DF for the variable and error, as well as the exact *P* value), to help readers interpret nonsignificant results.[6,19,22,29,36,45] Krzywinski and Altman[23] describe a nice comparison between ways of visualizing data variability. Reporting confidence intervals is particularly useful because it allows for post-hoc equivalence testing. This technique does allow one to differentiate true negative results from underpowered studies if the experimenter can give a meaningful value to the H_A. This value is typically the minimum difference you want to detect. Estimating a meaningful H_A is easier in clinical medicine, because the goal may be to detect a change from mild to moderate on a scale in which each interval is 10 points (for example[37]). If the minimal effect of interest falls outside of the 95% confidence interval generated by the equivalence test, then you had enough power. Smaller differences that fall within the confidence interval are unimportant, and one has grounds to accept H_0 and interpret the negative result as a true negative. In laboratory animal science we often use similar scales. For example, we might look back at old mouse nesting papers and say the typical change in nest scores is 1 point on the scale and would set that as our minimum difference to detect. Clinically, we might say a relevant change is 1 point in body condition scores or 1 point with the mouse grimace scale. Equivalence testing can easily be done with any statistical software.

Power as a performance standard

Researchers conduct experiments to understand biology and disease and have no desire to waste time, resources, or animal lives.[8] However, in a retrospective evaluation of power in behavioral ecology, only 10% to 20% of tests exceeded the recommended minimum criterion of 80% power.[19] In fact, some believe that we underestimate the number of low powered studies.[8,24] Due to this low power, Jennions and Moller[19] have encouraged researchers to increase sample sizes. While this recommendation is for a field other than laboratory animal science, it appears to be at odds with animal welfare and the 3Rs.[42] However, other approaches can be used to maximize statistical power while reducing animal numbers. Experimental design can be one of the most powerful tools for reducing animal use[8,10,11,16] but may not be applicable for all areas of research (such as case studies or nonexperimental surveys).

Factorial designs, in particular, maximize power.[16] When an experiment includes several treatment groups, the number of animals or cages in each group can be reduced. Although group sizes can be quite small (2 to 4 cages or animals per group), the experiment will always be more powerful than a single factor one.[10,11,13] For example, consider an experiment where a mouse model is being tested for the efficacy of a diet on sexual maturation (Figure 2). We might predict that the diet will accelerate sexual development, but this effect might differ between males and females. Such experiments are more-often-than-not incorrectly performed separately for each sex.[35] However, by analyzing the data together, you are able to control for known variation between the sexes,[25] even if this is not of particular interest. Furthermore, not only does a factorial experiment allow you to test these hypotheses with equal power using half the total number of animals, but it allows you test the secondary hypothesis— that the effect of diet '*differs*' between male and female animals (which is untestable if separate experiments are performed), and it allows positive- and negative-control post-hoc tests that are impossible otherwise.[35]

We present an example of such an experiment in Figure 2. In this example $n = 4$, but a total of 16 cages are used in the whole experiment ($N = 16$). It is important to note that '*n*' and '*N*' mean different things when the number of experimental units are being reported. *N* means the total number of experimental units that are used in the whole experiment, thus in our example $N = 16$ cages, not necessarily 16 animals. However, *n* means the number of experimental units in each combination of the treatments (each grayed cell in Figure 2). In our example, the smallest combination of treatments is $n = 4$. Reporting your experimental units as either *n* or *N* is equally correct but the terms convey slightly different information to the reader.

When using a general linear model (GLM) to test whether males and females have different ages of onset of sexual maturity (the positive control in this experiment), data from all 8 cages per sex (see row totals) are included. The same is true for the overall effect of the diet. This is partly where the hidden power of the factorial design comes from – a factorial experiment performs many experiments in one (it has an '*economy of variables*').[16] However, we are also interested in how the 2 sexes may react differently to the 2 diets, each combination of sex and diet would only have data from 4 cages (see grayed boxes in Figure 2).

To determine whether your experimental unit is the animal or perhaps a cage of animals, you must determine to what the treatment is being applied.[15] In this example, the diet is being applied to the cage of animals, and mice are housed in same-sex groups, so therefore the experimental unit is the cage. This is an important distinction that should be reported very clearly for readers and reviewers (Figure 2).

While factorial designs always improve power through an economy of variables, this is not their only benefit. Factorial designs are particularly effective when they are used to control for nuisance variation.[10,11,25] A nuisance variable causes variability in the data, which masks the real treatment effect.[10,11,25] Identifying these sources of variation, or nuisance variables, is critical during the planning stages.[8] Being able to isolate a factor that causes this variation (for example cell culture, PCR plate, or sex) by adding it to the statistical model increases sensitivity in detecting treatment effects and increases power.[10,11,25,30] These nuisance variables can be managed by using a randomized block design that incorporates the nuisance variables as 'blocks'. In mice, the home cage is a good example of a nuisance variable that can be used as a block

		Diet		
		Control	Drugged	Total
Sex	Males	4	4	8
	Females	4	4	8
	Total	8	8	16

Figure 2. Example 2 × 2 factorial design testing how diet may differentially affect the onset of sexual maturity due to a control and drugged diet.

to control variability from cage to cage.[11,15,49] Perhaps in a slightly different experimental design than the one previously proposed, we decide to inject the mice with the drug, instead of feeding it to them in the diet. Two of the 4 mice in the cage are injected with the test drug and other 2 with saline (a placebo that mimics handling effects but not the drug). In this new experiment the experimental unit would be the animal, cage is the blocking variable, and the design is still a factorial (each cage contains 2 drug-treated and 2 saline-treated animals). The power of this approach is that you know that animals within a block are more like each other than animals in different blocks (cages) but you don't need to know why or how the blocks differ (variability due to minute environmental differences between each cage, such as location on the rack). The block is included in the statistical model, and suitable analyses can now separate the variation between animals into between block (nuisance) error and between animal (measurement) error.[10,11,25] In more advanced designs, we can further subdivide into between-cage, between-animal, and measurement error—all without requiring any more animals.

An example can illustrate why this is so important. Imagine your primary outcome was body weight or some other variable with very little measurement error—the majority of the noise in your experiment is due to between-animal and between-cage variability. In a naïve experiment analyzed with a *t* test or a one-way ANOVA, these sources of error are confounded, error is over-estimated, and power is reduced.[25,26] The same data analyzed differently (see[15] for a flow chart to help decide which test is most appropriate) can require 2 to 10 times fewer animals to achieve similar power and significance.[25] The greater the influence of nuisance variables, the more this benefit accrues.[11,25]

In addition to increasing power, factorial designs can successfully mimic variation between independent replicates conducted within the same laboratory.[20,39,50] Heterogenization of study populations, by introducing variability in a controlled fashion, rather than arbitrarily trying to standardize everything, can be used to provide an estimate of the external validity and the potential reproducibility of results by systematically varying a few selected factors.[1,39,41] Any aspect of the animal, such as genotype, sex, age, body condition, or housing can be used to achieve systematic heterogenization.[1] In essence, this is planned and statistically controllable variation. In addition, this type of design more closely mimics human experimental designs where variability is embraced and controlled.[11-13] This simple use of powerful experimental designs that introduce systematic heterogenization is a proven step in improving reproducibility of animal-based experimentation.[11,12,39,40]

For more information about experimental designs, power, or designing animal based experiments, we highly recommend Festing and colleagues,[10] *The Design of Animal Experiments*. It is a very straight forward and approachable book on the statistics and design of animal experimentation. Dr Festing also has a free online interactive short course on experimental design that is highly recommended to anyone working with laboratory animals (http://3rs-reduction.co.uk/).

Conflict between IACUC requirements and powerful experimental designs

As part of the regulatory process, IACUC require an a priori justification for the requested number of animals. The intent is to make sure that applicants can justify the number of animals requested for their study. Typically, a power calculation is requested to assure that the investigators have thought through their experimental design and have reasonably estimated the number of animals needed to achieve sufficient power. A potential problem with this process is that power analyses for complicated experimental designs, such as factorial and randomized block designs, cannot be performed with simple formulae and can be impossibly daunting.[10,32] In fact, the study mentioned earlier[19] that evaluated the power of studies in behavioral ecology omitted these kinds of designs from their analysis because of this difficulty. However, a simple solution to calculate sample size for complicated designs is to use Mead resource equation.[10,33] This power estimate is suitable for complex biologic experiments, quantitative data, and any combination of categorical or continuous treatments and blocking variables.[10] It does not require an estimate of standard deviation or effect size, power, significance level, or alternative hypotheses.[10] Instead, it is based on curves of diminishing returns (see[11] for examples of how different designs and blocking can changes these curves) as the sample size increases. As you add more animals, you receive less and less benefit in terms of the critical value of the *F*-test and on the accuracy of estimating the variance components in general linear and mixed models.[10,33] Thus, Mead recommends that the error DF fall between 10 and 20 (see the worked example below). There are 2 sides to the beautiful simplicity of this reasoning. First, once the error DF for an estimated *F*-ratio has reached 20, there are little to no marginal gains in terms of the accuracy of the estimate. In other words, an effect that is not detectable by 20 error DF is probably not going to be worth the effort of trying to detect with larger sample sizes. Second, different sources of variance behave differently with increasing sample size—for instance in a randomized block design, treatment mean-squares increase linearly with sample size, while block and residual (error) mean-squares stabilize at a constant value (which is typically achieved around 20 error DF). As a result, the greater the between-block variance, the more Mead equation holds true. Furthermore, the equation holds true for a wide range of effect sizes—as long as nuisance variables are included, and between-block variance outweighs measurement error, effect size has little bearing on LSN which is where the 10×, 100×, 1000× sample size cost of simple *t* tests arises.[11]

A number of standard equations can be used to calculate error DF, but the simplest solution is to simulate data (given expected effect sizes for both treatment and blocking factors), run the appropriate analysis, and check the corresponding error DF from the statistical output. In fact, we strongly recommend that simulated data and analysis be part of a number justification. We realize that simulating data may be difficult for many and at the very least we suggest providing a simplified ANOVA table. This should include DF for the various terms and treatments in the analysis similar to what we illustrate below.

Continuing with the 2 × 2 factorial examples from earlier (Figure 2), we need to determine if 16 total cages will be sufficient for this study based on Mead resource equation.[10,33] While estimating the number of cages may not be as relevant to the IACUC as total animal numbers, this factor still needs to be determined, and the experimental unit defined as part of the animal numbers justification. In order to calculate DF you take the items value minus 1. The Total DF (or total number of experimental units or observations for our experiment) = 16 - 1 =15; the sex DF (we have 2 levels for male and female) = 2 - 1 = 1; the diet DF (there are 2 treatments drug and control) = 2 - 1 = 1; and the sex-by-diet interaction DF (multiply the DF for each variable) = 1 × 1 = 1. Therefore, the total DF for our statistical model = 1+1+1= 3; and the error DF = 15(total DF) - 3(model DF) = 12. Twelve falls between 10 and 20 and reducing the number of cages per combination from $n = 4$ ($N = 16$) to $n = 3$ ($N = 12$) would result in 8 error DF, which would be below our minimum of 10 DF. Increasing the number of cages from $n = 4$ ($N = 16$) to $n = 6$ ($N = 24$) would result in 20 error DF. We recommend the '*Price is Right*' game show philosophy, (that is, pick the closest value to 20 without going over) when considering your sample size. Therefore in this situation, we'd recommend that $n = 6$ cages ($N = 24$). While we have established the number of cages needed for this experiment, the total number of mice has yet to be determined. Depending on what kind of conditions the researcher wishes to mimic – for example a group housing scenario or single housing—will affect the total number of animals but not the power of the experiment. Only one data point can be recorded for a cage, regardless of whether there were 1 or 5 mice in each cage. Thus, a measure of body weight for a cage of 5 mice would be averaged across the cage for a group housing situation.

In the case of the randomized block design we considered (where 2 animals per cage each receive an injection of either saline or drug; mouse is the experimental unit): Drug requires 1 DF; cage is the block, but is nested within sex (each cage only contains one mouse sex). There is 1 DF for sex. So to calculate the block with 16 cages, we have (16 - 1)-1 (for the sex DF) =14 cage block DF; and 1 DF for the sex-by-drug interaction (1 × 1 = 1). Therefore, the model uses 1 (sex) +14 (block)+ 1 (drug) + 1 (sex*drug) = 17 DF. We have 16 cages*4 mice per cage = 64 mice in total ($N = 64$), and so we have 63 total DF. The error DF is therefore 63 - 17 = 46, and we see that the experiment is actually overpowered (that is, greater than 20). In fact, we would only need a total of 6 cages ($N = 24$ mice; Total DF = 23; Model DF = 1+4+1+1 = 7; Error DF = 16).

In terms of the 3Rs, simply taking the time to consider a slightly different experimental design and statistical test can markedly REDUCE animal use, as shown by our examples above, in which we reduced the number of mice by 40 (that is, to a third of the original number) while retaining statistical power.

Conclusion

Statistics and experimental design are essential to the process of conducting sufficiently powered research, but equally important in helping reduce the number of animals used. Before anything else, researchers should first decide what design and which statistical test will be used to analyze their data. These choices inherently influence the power of the subsequent study, whether it can be calculated, and the number of animals needed. Only after making these decisions can sample size be estimated. Although formal power calculations can be done for statistical tests such as one-way ANOVAs, we would like to encourage the use of a more broad application of sophisticated experimental designs (such as randomized block designs) to increase power while simultaneously reducing animal numbers. Incorporating nuisance variables and systematic heterogeneity will inevitably help reduce data noise and improve reproducibility.[11,12,39,40] However, depending on the researcher's decisions about their design and choice of statistical test, sample size may be too complicated to calculate with more complex designs. Mead resource equation can be used to simply and appropriately estimate sample sizes for more complex factorial designs while maintaining an acceptable amount of statistical power. While we believe that traditional power calculations can still be useful in certain circumstances, we believe there are other designs and methods available that better address the 3Rs.

Two important take-home messages are that 1) even well-powered experiments can produce null results and 2) estimating post-hoc power from a completed experiment is not a valid way of determining whether the result is a true one. The only appropriate way to determine this is to use equivalence testing or at a minimum to calculate the LSN of experimental units that would be necessary to produce sufficient power. That way an ethical decision can be made to determine whether substantial animal use (for example, achieving 80% power will require the use of 500 animals) is worth the result. To further improve interpretation of results, we recommend that authors report confidence intervals, observed effect sizes, and test statistics, especially for null results, so that readers and reviewers can make an educated assessment of the result.

Many researchers may not feel confident of their knowledge when it comes to statistics and experimental design. In that case, we would highly recommend they consult a biostatistician who can help them navigate these ideas and suggestions early in the planning process. The creation of high-quality studies is best achieved with collaborative discussions between the researcher and experts in other disciplines such as statistics, laboratory animal science, and comparative medicine.[8,9] The simple act of asking for help has the potential to improve the translation of animal-based experiments and reduce the number of animals needed to complete them.

Acknowledgments

Thank you to Kathleen Pritchett-Corning and Linda Toth for reviewing the manuscript and helping us further clarify our message and reduce jargon.

References

1. **Bailoo JD, Reichlin TS, Würbel H.** 2014. Refinement of experimental design and conduct in laboratory animal research. ILAR J **55:**383–391. https://doi.org/10.1093/ilar/ilu037.
2. **Begley CG.** 2013. Six red flags for suspect work. Nature **497:**433–434. https://doi.org/10.1038/497433a.
3. **Begley CG, Ellis LM.** 2012. Drug development: raise standards for preclinical cancer research. Nature **483:**531–533. https://doi.org/10.1038/483531a.
4. **Blackshaw JK, Swain AJ, Blackshaw AW, Thomas FJM, Gillies KJ.** 1997. The development of playful behaviour in piglets from birth to weaning in three farrowing environments. Appl Anim Behav Sci **55:**37–49. https://doi.org/10.1016/S0168-1591(97)00034-8.
5. **Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, Munafò MR.** 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci **14:**365–376. https://doi.org/10.1038/nrn3475.
6. **Coe R**. [Internet]. 2002. It's the effect size, stupid: What effect size is and why it is important. Presented at the Annual Conference of the British Educational Research Association. University of Exeter,

England. 12–14 September 2002. [Cited 15 August 2019]. Available at: http://www.leeds.ac.uk/educol/documents/00002182.htm
7. **Cohen J.** 1969. Statistical power analysis for the behavioral sciences, 2nd ed. New York (NY): Academic Press.
8. **Everitt JI.** 2015. The future of preclinical animal models in pharmaceutical discovery and development: a need to bring in cerebro to the in vivo discussions. Toxicol Pathol **43:**70–77. https://doi.org/10.1177/0192623314555162.
9. **Everitt JI, Berridge BR.** 2017. The role of the IACUC in the design and conduct of animal experiments that contribute to translational success. ILAR J **58:**129–134. https://doi.org/10.1093/ilar/ilx003.
10. **Festing MF, Overend P, Gaines Das R, Cortina-Borja M, Berdoy M.** 2002. The design of animal experiments: reducing the use of animals in research through better experimental design. London (United Kingdom): The Royal Society of Medicine Press Limited.
11. **Garner JP.** 2014. The significance of meaning: why do over 90% of behavioral neuroscience results fail to translate to humans, and what can we do to fix it? ILAR J **55:**438–456. https://doi.org/10.1093/ilar/ilu047.
12. **Garner JP, Gaskill BN, Weber EM, Ahloy-Dallaire J, Pritchett-Corning KR.** 2017. Introducing Therioepistemology: the study of how knowledge is gained from animal research. Lab Anim (NY) **46:**103–113. https://doi.org/10.1038/laban.1224.
13. **Gaskill BN, Stottler AM, Garner JP, Winnicker CW, Mulder GB, Pritchett-Corning KR.** 2017. The effect of early life experience, environment, and genetic factors on spontaneous home-cage aggression-related wounding in male C57BL/6 mice. Lab Anim (NY) **46:**176–184. https://doi.org/10.1038/laban.1225.Erratum in: Lab Anim (NY) 2019.
14. **Goodman SN, Berlin JA.** 1994. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. Ann Intern Med **121:**200–206. https://doi.org/10.7326/0003-4819-121-3-199408010-00008.
15. **Gosselin RD.** 2019. Guidelines on statistics for researchers using laboratory animals: the essentials. Lab Anim **53:**28–42. https://doi.org/10.1177/0023677218783223.
16. **Grafen A, Hails R.** 2002. Modern statistics for the life sciences. Oxford (United Kingdom): Oxford University Press.
17. **Hoenig JM, Heisey DM.** 2001. The abuse of power: The pervasive fallacy of power calculations for data analysis. Am Stat **55:**19–24. https://doi.org/10.1198/000313001300339897.
18. **Ioannidis JP.** 2005. Why most published research findings are false. PLoS Med **2:**0696–0701. https://doi.org/10.1371/journal.pmed.0020124.
19. **Jennions MD, Møller AP.** 2003. A survey of the statistical power of research in behavioral ecology and animal behavior. Behav Ecol **14:**438–445. https://doi.org/10.1093/beheco/14.3.438.
20. **Jonker RM, Guenther A, Engqvist L, Schmoll T.** 2013. Does systematic variation improve the reproducibility of animal experiments? Nat Methods **10:**373. https://doi.org/10.1038/nmeth.2439.
21. **Kilkenny C, Browne WJ, Cuthill IC, Emerson M, Altman DG.** 2010. Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. PLoS Biol **8:**1–5. https://doi.org/10.1371/journal.pbio.1000412.
22. **Kraemer HC, Mintz J, Noda A, Tinklenberg J, Yesavage JA.** 2006. Caution regarding the use of pilot studies to guide power calculations for study proposals. Arch Gen Psychiatry **63:**484–489. https://doi.org/10.1001/archpsyc.63.5.484.
23. **Krzywinski M, Altman N.** 2013. Error bars. Nat Methods **10:**921–922. https://doi.org/10.1038/nmeth.2659.
24. **Krzywinski M, Altman N.** 2013. Power and sample size. Nat Methods **10:**1139–1140. https://doi.org/10.1038/nmeth.2738. Erratum was publishing in 2014 and 2015.
25. **Krzywinski M, Altman N.** 2014. Analysis of variance and blocking. Nat Methods **11:**699–700. https://doi.org/10.1038/nmeth.3005.
26. **Krzywinski M, Altman N, Blainey P.** 2014. Nested designs. Nat Methods **11:**977–978. https://doi.org/10.1038/nmeth.3137.
27. **Lenhard W, Lenhard A**. 2015. Calculation of effect sizes. Dettelbach (Germany): Psychometrica.
28. **Leung V, Rousseau-Blass F, Beauchamp G, Pang DSJ.** 2018. ARRIVE has not ARRIVEd: Support for the ARRIVE (Animal Research: Reporting of in vivo Experiments) guidelines does not improve the reporting quality of papers in animal welfare, analgesia or anesthesia. PLoS One **13:**1–13. https://doi.org/10.1371/journal.pone.0197882.
29. **Levine M, Ensom MH.** 2001. Post hoc power analysis: An idea whose time has passed? Pharmacotherapy **21:**405–409. https://doi.org/10.1592/phco.21.5.405.34503.
30. **Lossie AC, Lo CL, Baumgarner KM, Cramer MJ, Garner JP, Justice MJ.** 2012. ENU mutagenesis reveals that *Notchless homolog* 1 (*Drosophila*) affects *Cdkn1a* and several members of the *Wnt* pathway during murine pre-implantation development. BMC Genet **13:**1–16. https://doi.org/10.1186/1471-2156-13-106.
31. **Macleod MR.** [Internet]. 2017. Findings of a retrospective, controlled cohort study of the impact of a change in Nature journals' editorial policy for life sciences research on the completeness of reporting study design and execution. bioRxivorg [Cited 24 July 2018]. Available at: https://www.biorxiv.org/content/10.1101/187245v1
32. **Marino MJ.** 2014. The use and misuse of statistical methodologies in pharmacology research. Biochem Pharmacol **87:**78–92. https://doi.org/10.1016/j.bcp.2013.05.017.
33. **Mead R.** 1988. The design of experiments: statistical principles for practical applications. New York (NY): Cambridge University Press.
34. **Nakagawa S, Cuthill IC.** 2007. Effect size, confidence interval and statistical significance: a practical guide for biologists. Biol Rev Camb Philos Soc **82:**591–605. https://doi.org/10.1111/j.1469-185X.2007.00027.x. Erratum: Biol Rev Camb Philos Soc 2009.**84:**515.
35. **Nieuwenhuis S, Forstmann BU, Wagenmakers E-J.** 2011. Erroneous analyses of interactions in neuroscience: a problem of significance. Nat Neurosci **14:**1105–1107. https://doi.org/10.1038/nn.2886.
36. **O'Keefe DJ.** 2007. Brief report: Post hoc power, observed power, a priori power, retrospective power, prospective power, achieved power: sorting out appropriate uses of statistical power analyses. Commun Methods Meas **1:**291–299. https://doi.org/10.1080/19312450701641375.
37. **Parker KJ, Oztan O, Libove RA, Mohsin N, Karhson DS, Sumiyoshi RD, Summers JE, Hinman KE, Motonaga KS, Phillips JM**. 2019. A randomized placebo-controlled pilot trial shows that intranasal vasopressin improves social deficits in children with autism. Sci Transl Med **11:**eaau7356. https://doi.org/10.1126/scitranslmed.aau7356
38. **Richardson JTE.** 2011. Eta squared and partial eta squared as measures of effect size in educational research. Educational Research Review **6:**135–147. https://doi.org/10.1016/j.edurev.2010.12.001.
39. **Richter SH, Garner JP, Auer C, Kunert J, Würbel H.** 2010. Systematic variation improves reproducibility of animal experiments. Nat Methods **7:**167–168. https://doi.org/10.1038/nmeth0310-167.
40. **Richter SH, Garner JP, Würbel H.** 2009. Environmental standardization: cure or cause of poor reproducibility in animal experiments? Nat Methods **6:**257–261. https://doi.org/10.1038/nmeth.1312.
41. **Richter SH, Garner JP, Zipser B, Lewejohann L, Sachser N, Touma C, Schindler B, Chourbaji S, Brandwein C, Gass P, van Stipdonk N, van der Harst J, Spruijt B, Võikar V, Wolfer DP, Würbel H.** 2011. Effect of population heterogenization on the reproducibility of mouse behavior: a multi-laboratory study. PLoS One **6:**1–14. https://doi.org/10.1371/journal.pone.0016461.
42. **Russell WMS, Burch RL.** 1959. The principles of humane experimental technique. London (United Kingdom): Methuen.
43. **Shanks DR, Vadillo MA, Riedel B, Clymo A, Govind S, Hickin N, Tamman AJ, Puhlmann LM.** 2015. Romance, risk, and replication: Can consumer choices and risk-taking be primed by mating motives? J Exp Psychol Gen **144:**e142–e158. https://doi.org/10.1037/xge0000116.
44. **Sullivan GM, Feinn R.** 2012. Using effect size-or why the *P* value is not enough. J Grad Med Educ **4:**279–282. https://doi.org/10.4300/JGME-D-12-00156.1.
45. **Sun S, Pan W, Wang LL.** 2011. Rethinking observed power: Concept, practice, and implications. Methodology (Gott) **7:**81–87.
46. **Thompson B**. [Internet]. 1999. Common methodology mistakes in educational research, revisited, along with a primer on both

effect sizes and the bootstrap. [Cited 15 August 2019]. Available at: https://eric.ed.gov/?id=ED429110

47. **Toth LA, Duffee NE.** 2017. Publication of negative data contributes to sound science. J Am Assoc Lab Anim Sci **56:**487–487.
48. **Würbel H.** 2000. Behaviour and the standardization fallacy. Nat Genet **26:**263. https://doi.org/10.1038/81541.
49. **Würbel H, Garner JP**. 2007. Refinement of rodent research through environmental enrichment and systematic randomizations. NC3Rs **9:**1–9.
50. **Würbel H, Richter SH, Garner JP.** 2013. Reply to: "Reanalysis of Richter et al.(2010) on reproducibility". Nat Methods **10:**374. https://doi.org/10.1038/nmeth.2446.