Simulation-Based Power-Analysis for Factorial ANOVA Designs

Daniël Lakens[1] & Aaron R. Caldwell[2,3]

[1] Human-Technology Interaction Group, Eindhoven University of Technology, The
Netherlands

[2] Department of Health, Human Performance and Recreation, University of Arkansas, USA

[3] Thermal and Mountain Medicine Division, U.S. Army Research Institute of Environmental
Medicine, USA

Author Note

Correspondence concerning this article should be addressed to Daniël Lakens, ATLAS
9.402, 5600 MB, Eindhoven, The Netherlands. E-mail: D.Lakens@tue.nl

Abstract

Researchers often rely on analysis of variance (ANOVA) when they report results of experiments. To ensure a study is adequately powered to yield informative results when performing an ANOVA, researchers can perform an a-priori power analysis. However, power analysis for factorial ANOVA designs is often a challenge. Current software solutions do not enable power analyses for complex designs with several within-subject factors. Moreover, power analyses often need partial eta-squared or Cohen's f as input, but these effect sizes are not intuitive and do not generalize to different experimental designs. We have created the R package Superpower and an online Shiny app to enable researchers without extensive programming experience to perform simulation-based power analysis for ANOVA designs of up to three within- or between-subject factors, with an unlimited number of levels. Predicted effects are entered by specifying means, standard deviations, and correlations (for within-subject factors). The simulation provides the statistical power for all ANOVA main effects, interactions, and individual comparisons, and allow researchers to correct for multiple comparisons. The simulation plots p-value distributions for all tests, and power plots across a range of sample sizes. This tutorial will demonstrate how to perform power analysis for ANOVA designs, and highlights important factors that determine the statistical power of factorial ANOVA designs.

Keywords: power analysis, ANOVA, hypothesis test, sample size justification, repeated measures

Word count: 4654 words.

Simulation-Based Power-Analysis for Factorial ANOVA Designs

When a researcher aims to test hypotheses with an analysis of variance (ANOVA), the sample size of the study should be justified based on the statistical power of the test. The statistical power of a test is the probability of rejecting the null-hypothesis, given a specified effect size, alpha level, and sample size. When power is low there is a high probability of concluding there is no effect when an underlying effect may exist in the population of interest. Several excellent resources exist that explain power analyses, including books (Aberson, 2019; Cohen, 1988), general reviews (Maxwell, Kelley, & Rausch, 2008), and practical primers (Brysbaert, 2019; Perugini, Gallucci, & Costantini, 2018). Whereas power analyses for individual comparisons are relatively easy to perform, power analyses for factorial ANOVA designs are a bigger challenge. There are many current software solutions that *can* calculate power for factorial ANOVAs (Campbell & Thompson, 2012; Faul, Erdfelder, Lang, & Buchner, 2007; Lang, 2017; J. Westfall, 2015a), but the options within these packages are often limited (e.g, Gpower has limited options for within subjects factors). Available software solutions do not provide easy options to specify more complex designs (e.g., a 2x2x2 design, where the first factor is manipulated between participants, and the last two factors are manipulated within participants). The predicted effects often need to be specified as a standardized effect size such as Cohen's f or partial eta squared (η_p^2), which are not the most intuitive way to specify a hypothesized pattern of results, and these effect sizes do not generalize to different experimental designs. Simulations based on a specified pattern of means and a covariance matrix (based on the expected standard deviation and correlation between within participant factors) provide a more flexible approach to power analyses. However, such simulations typically require extensive programming knowledge.

In this manuscript we introduce Superpower, an R package and Shiny app that can be used to perform power analyses for factorial ANOVA designs based on simulations. Superpower can be used to calculate the statistical power based on a predicted pattern of

means, standard deviations, and correlations (for within-subject factors). By simulating data for factorial designs with specific parameters researchers can gain a better understanding of the factors that determine the statistical power of an ANOVA, and learn how to design well-powered experiments. After a short introduction to statistical power, focussing on the F-test, we will illustrate through simulations how the power of factorial ANOVA designs depend on the pattern of means across conditions, the number of factors and levels, the sample size, and whether you need to control the alpha level for multiple comparisons.

A simple scenario for power analysis

Imagine you plan to perform a study in which participants interact with an artificial voice assistant who sounds either cheerful or sad. You measure how much 80 participants in each condition enjoy interacting with with the voice assistant on a line marking scale (coded continuously from -5 to 5) and observe a mean of 0 in the sad condition, and a mean of 1 in the cheerful condition, with an estimated standard deviation of 2.

After submitting your manuscript for publication, reviewers ask you to add a study with a neutral control condition to examine whether cheerful voices increase, or sad voices decrease enjoyment (or both). Depending on what the mean enjoyment in the neutral condition is, what sample size would you need to collect for a high powered test of the expected pattern of means? A collaborator suggests to switch from a between-subject design to a within-subject design to collect data more efficiently. What impact will switching from a between-subject to a within-subject design have on the required sample size? The effect size in the first study could be considered "medium" based on the benchmarks by Cohen (1988), but does it make sense to plan for a "medium" effect size in either the between-subject or within-subject ANOVA design? And if you justify the sample size based on the power for the main effect for the ANOVA, will the study also have sufficient statistical power for the independent comparisons between conditions (or vice versa)?

Calculating Power for ANOVAs

Before we directly answer these questions, let's consider some of the basic concepts of statistical power and how power calculations are typically performed. Let's consider the initial study described above, where enjoyment is measured when 80 participants per condition interact with a cheerful or sad voice assistant. We can test the difference between two means with a t-test or a one-way ANOVA, and the two tests are mathematically equivalent. Figure 1 and Figure 2 visualize the distribution of the effect sizes Cohen's d (for the t-test) and η_p^2 (for the F-test) that should be observed when there is no effect (grey curves) and when the observed difference between means equals the true effect (black curves). In both figures the light grey areas under the null-distribution mark the observed results that would lead to a Type 1 error (observing a statistically significant result if the null-hypothesis is true) and the dark grey areas under the curve mark the observed effect sizes that would lead to a Type 2 error (observing a non-significant result when there is a true effect).

When the observed data are more extreme than a critical value, the test result is statistically significant. Critical values are often expressed as t-values or F-values, but for a given sample size they can also be expressed as critical *effect sizes*. Given the sample size of 80 participants per group, observed effects are statistically significant when they are larger than d = 0.31 in a t-test, or $\eta_p^2 = 0.02$ for the F-test. The goal of an a-priori power analysis is to choose a sample size for which the probability of observing a statistically significant effect reaches a desired probability. To calculate power, one has to specify the alternative hypothesis (the black curves in Figure 1 and 2). If we assume that under the alternative hypothesis the true effect size is d = 0.5 or $\eta_p^2 = 0.0588$, and data are collected from 80 participants in each condition, in the long run 8,816.02% of the observed data will yield statistically significant results.

The t-test can examine the difference between two means $(m_1 - m_2)$, and the F-test computes the ratio of the between group variance and the within group variance. For two
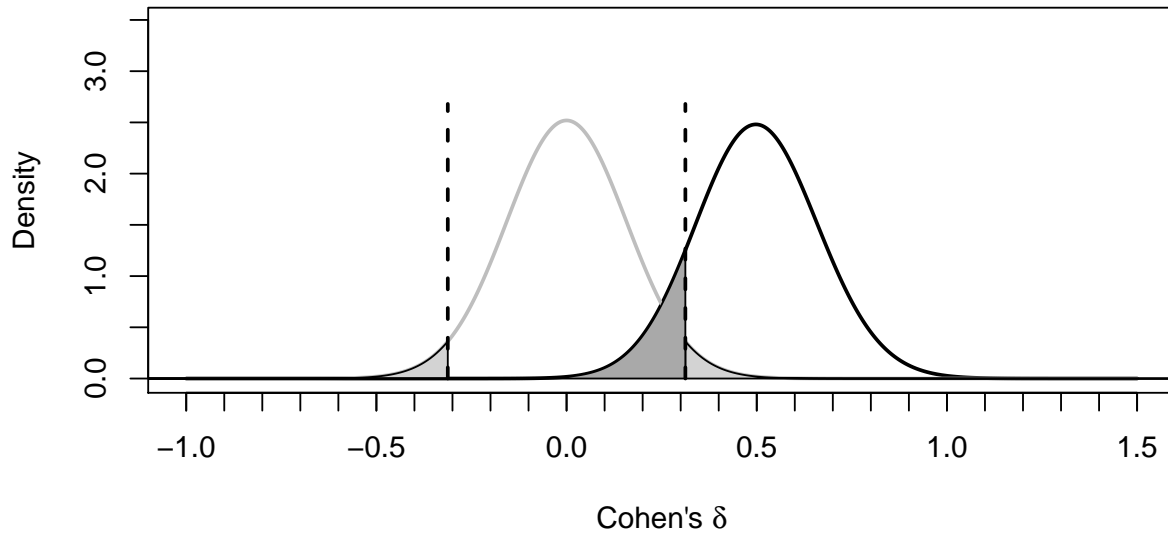
Figure 1. Distribution of Cohen's d under the null-hypothesis (grey curve) and alternative hypothesis assuming d = 0.5 (black curve).

independent groups, the *t*-statistic can easily be translated to the *F*-ratio $F = t^2$. Cohen's d, a standardized effect size, is calculated by dividing the difference between sample means by the standard deviation, or

$$d = \frac{m_1 - m_2}{\sigma}. \tag{1}$$

If we have two groups with the sample means of 0 and 1, and the standard deviation is 2, Cohen's d is (1-0)/2, or 0.5. The generalization of Cohen's d to more than two groups is Cohen's f, which is the standard deviation of the population means divided by the population standard deviation (Cohen, 1988), or:

$$f = \frac{\sigma_m}{\sigma} \tag{2}$$

where for equal sample sizes,

$$\sigma_m = \sqrt{\frac{\sum_{i=1}^{k}(m_i - m)^2}{k}}. \tag{3}$$

For two groups Cohen's f is half as large as Cohen's d, or $f = \frac{1}{2}d$. Because Cohen's f is the
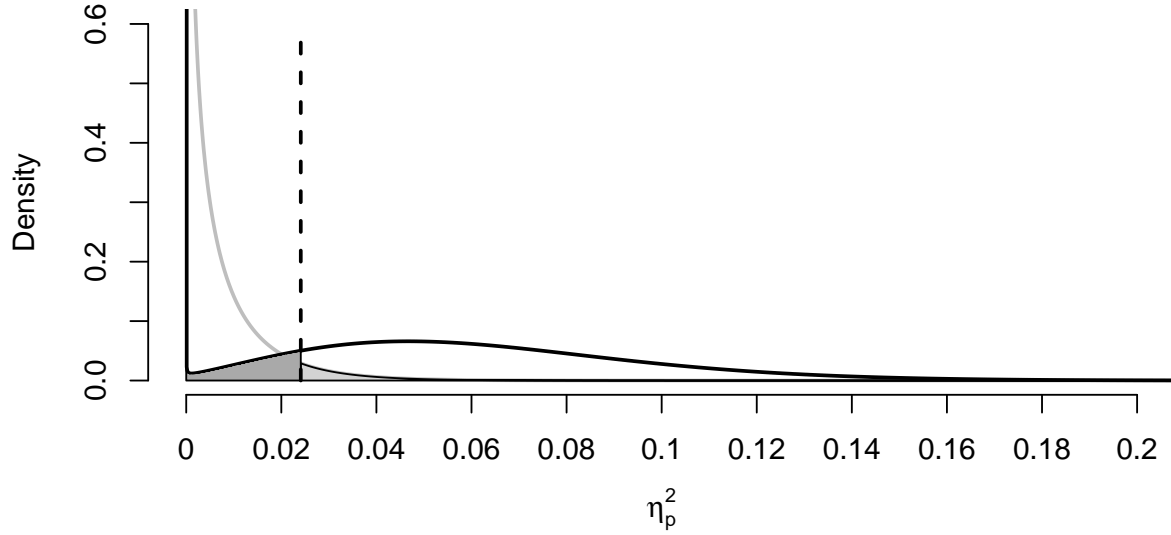
Figure 2. Distribution of eta-squared under the null-hypothesis (grey curve) and alternative hypothesis assuming partial eta-squared = 0.0588 (black curve).

effect size used in power analyses for ANOVA designs it is worth illustrating how Cohen's f is calculated. If we again take two means of 1 and 2, and a standard deviation of 2, the grand mean is 1.5. We subtract each condition mean from the grand mean, square it, calculate the sum of squares, divide it by two, and take the square root.
$\sigma_m = \sqrt{\frac{(0-0.5)^2+(1-0.5)^2}{2}} = \sqrt{\frac{0.25+0.25}{2}} = 0.5$. Dividing this value by the standard deviation yields $f = \frac{0.5}{2} = 0.25$.

G*Power (Faul et al., 2007), a very popular program for calculating sample size planning and power among social scientists, allows researchers to enter the effect size for the power analysis for ANOVA designs as either Cohen's f or partial eta-squared (η_p^2). Partial eta-squared can be converted into Cohen's f:

$$f = \sqrt{\frac{\eta_p^2}{1 - \eta_p^2}} \tag{4}$$

and Cohen's f can be converted into partial eta-squared:

$$\eta_p^2 = \sqrt{\frac{f^2}{f^2 + 1}} \tag{5}$$

In the example above, $\eta_p^2 = 0.25^2/(0.25^2 + 1) = 0.0588$.

Power calculations rely on the noncentrality parameter (lambda, (λ).) In a between-participants one-way ANOVA lambda is calculated as:

$$\lambda = f^2 \times N \tag{6}$$

where f is Cohen's f and N is the total sample size. Based on λ (which is the noncentrality parameter of the noncentral F distribution, Figure 2) and the critical test statistic (which specifies the part of the distribution that is more extreme than the test statistic needed for a statistically significant test result) we can calculate how much of the distribution under the alternative hypothesis will be statistically significant in the long run (i.e., the area under the black curve in Figure 2 to the right of the critical effect size).

Simulating Statistical Power for Different Factorial Designs

The code underlying the Superpower R package and the Shiny app generates data for each condition in the design and performs an ANOVA and *t*-tests for all comparisons between conditions. The percentage of significant results (i.e., the power) is calculated based on the desired alpha level, and average effect size estimates are computed for all tests. Users need to specify the design of the study by providing the number of levels for each factor and indicating whether the factor is measured between or within participants. We can create such a design using Superpower's ANOVA_design function.

```
design_result <- ANOVA_design(design = "2b", n = 80,

                              mu = c(1, 0), sd = 2,

                              labelnames = c("condition",

                                             "cheerful", "sad"))
```

Our initial study above is a "2b" or two level between-participant design. For ease of interpretation the factors and levels can be named (for our example, factor = "Condition" with levels = "cheerful" and "sad"). The number of observations per condition should be specified (i.e, 80 participants in each between-participant condition). The means for each condition should be entered (0 and 1), as well as the standard deviation (2). For within-participant designs the correlation between variables should be specified. A document detailing how to reproduce all simulations in this manuscript using Superpower of the Shiny app is available in the appendix.
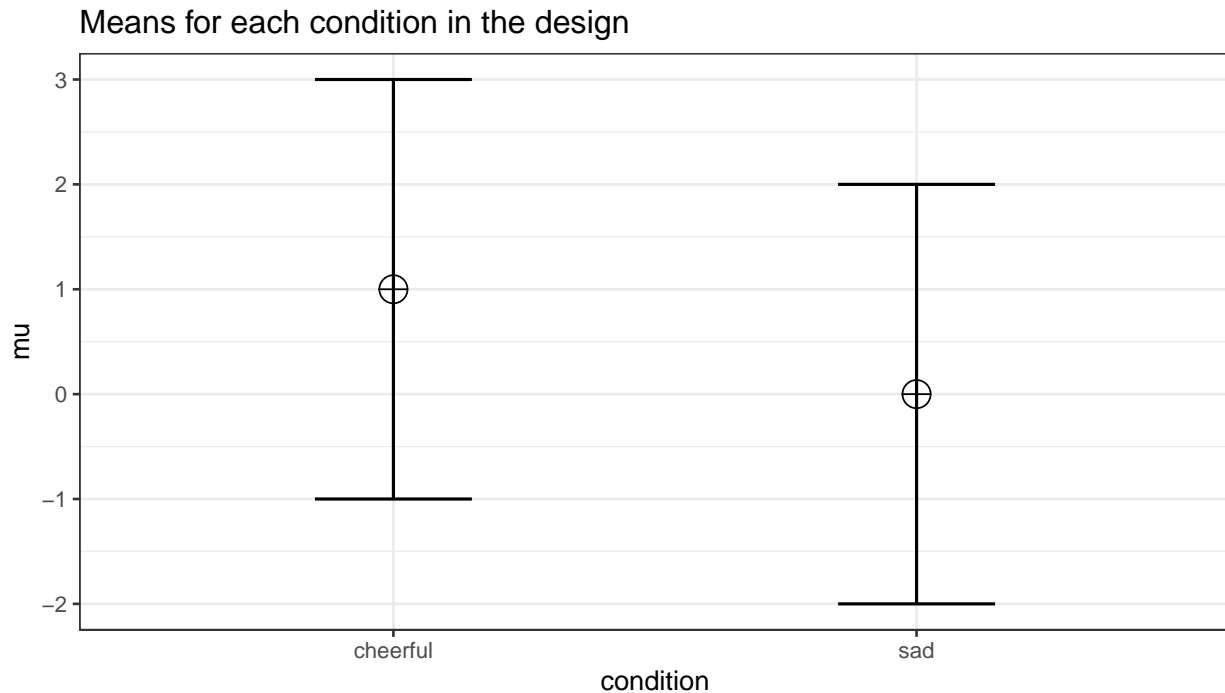
Means for each condition in the design

Figure 3. Vizualization for the means and standard deviations for the two conditions. Error bars represent one standard deviation.

For a visual confirmation of the input, a figure is created that displays the means and standard deviation (see Figure 3).

There are two ways to calculate the statistical power of an ANOVA through simulations. The first is to repeatedly simulate datasets and compute the percentage of

statistically significant results. This is completed with the ANOVA_power function which performs a Monte Carlo simulation.

```
power_result_monte <- ANOVA_power(design_result, nsims = 100000)
```

The second is to simulate a dataset that has *exactly* the desired properties, perform an ANOVA, and use the ANOVA results to compute the statistical power. This is completed with the `ANOVA_exact` function which performs a "exact" simulation.

```
power_result_exact <- ANOVA_exact(design_result)
```

The first approach is a bit more flexible, but the second approach is much faster. Superpower let's users choose either of these simulation approaches. Because the true pattern of the data are unknown in an a-priori power analysis, there is often uncertainty about the values that need to be entered in a power analysis. It makes sense to examine power across a range of assumptions, from more optimistic scenarios, to more conservative estimates. It is recommended to power not for the pattern of means you expect, but the smallest effect size that you consider worthwhile to detect (for examples, see Lakens, Scheel, and Isager (2018)). The results of a simulation study will vary each time the simulation is performed (but can be made reproducible by specifying a "seed" number, e.g., `set.seed = 2019`). To perform the simulations, you specify the number of simulations, the alpha level for the tests, and any adjustments for multiple comparisons that are required. If ANOVA_power is being used, the more simulations that are performed, the more stable the results, but the longer the simulation will take.

If 100.000 simulations are performed for our two group between subjects design with means of 1 and 0, a standard deviation of 2, and 80 participants in each group, with a seed set to 2019 (these settings will be used for all simulation results reported in this manuscript), the statistical power (based on the percentage of $p < \alpha$ results) is 88.19% and the average η_p^2

is 0.06. The simulation also provides the results for the individual comparisons based on
t-tests. Since there are only two groups in this example, the statistical power for the
individual comparison is identical to the ANOVA, but the expected effect size, Cohen's d, is
-0.50.

Now that the basic idea behind the simulation is clear, we can use simulations to
explore how changes to the experimental design influence power, and answer some of the
questions our hypothetical researcher is confronted with when designing a follow-up study.
We will first examine what happens if we add a third, neutral, condition to the design. Let's
assume the mean enjoyment rating for the neutral voice condition falls either perfectly
between the cheerful and sad conditions, or is equal to the cheerful condition. Again, we set
this with ANOVA_design.

```
design_result_1 <- ANOVA_design(design = "3b", n = 80,
                                mu = c(1, 0.5, 0),
                                sd = 2,
                                labelnames = c("condition",
                                               "cheerful",
                                               "neutral", "sad"))
```

The design now has 3 between-participant conditions, and we can explore what
happens if we would collect 80 participants in each condition.

If we assume the mean falls exactly between between the cheerful and sad conditions,
the simulations show the statistical power for a 3-groups one-way ANOVA *F*-test is reduced
to 81.14 %, and the effect size (partial eta-squared) is 0.05. If we assume the mean is equal
to the cheerful condition, the power increases to 91.03v%. Compared to the two group
design (where the power was 88.19%), three things have changed. First, the numerator
degrees of freedom has increased because an additional group is added to the design, which
makes the non-central *F*-distribution more similar to the central *F*-distribution, which

reduces the statistical power. Second, the total sample size is 50% larger after adding 80 participants in the third condition, which increases the statistical power of the ANOVA. Third, the effect size, Cohen's f, has decreased from 0.25 to either 0.20 or 0.24, which reduces the statistical power. The exact effect of these three changes on the statistical power is difficult to anticipate from one ANOVA design to the next, but simulations can help illustrate the effects of these changes. The most important conclusion based on these simulations is that changing an experimental design can have several opposing effects on the power of a study, depending of the pattern of means. One cannot assume the effect size remains unchanged when the design changes, and it might make more sense to think about expected effects in terms of the pattern of means and variance across conditions.

Power for Within-Subject Designs

What happens if we would perform the second study as a within-participants design? Instead of collecting three groups of participants, we only collect one group, and let this group evaluate the cheerful, neutral, and sad voice assistants. Again, we set this up with the ANOVA_design function and notice the only differences are changing 3b to 3w and the correlation parameter, r = 0.5 is specificied within the function.

```
design_result_within_1 <- ANOVA_design(
  design = "3w",
  n = 80,
  mu = c(1, 0.5, 0),
  sd = 2,
  r = 0.5,
  labelnames = c("condition", "cheerful", "neutral", "sad")
)
```

A rough but useful approximation of the sample size needed in a within-subject design

(N_W), relative to the sample needed in between-design (N_B), is (from Maxwell & Delaney, 2004, p. 562, formula 47):

$$N_W = \frac{N_B(1 - \rho)}{a} \qquad (7)$$

Here a is the number of within-participant levels, ρ is the correlation between the measurements. Even if the correlation is 0, the required sample size for a within design is reduced because each participants contributes multiple measurements. A positive correlation reduces the standard deviation of the difference scores and increases the statistical power. Because the standardized effect size is the mean difference divided by the standard deviation of the difference scores, the correlation has an effect on the standardized mean difference in a within-subject design, referred to as Cohen's d_z (because it is the effect size of the difference score between x and y, z). The relation is:

$$\sigma_z = \sigma\sqrt{2(1 - \rho)} \qquad (8)$$

Cohen's d_z is used in power analyses for dependent t-tests, but there is no equivalent Cohen's f_z for a within-participant ANOVA, and Cohen's f is identical for within and between designs. Instead, the value for lambda (λ) is adjusted based on the correlation. For a one-way within-participant design lambda is identical to Equation (6), multiplied by u, a correction for within-subject designs, calculated as:

$$u = \frac{k}{1 - \rho} \qquad (9)$$

where k is the number of levels of the within-participant factor, and ρ is the correlation between dependent variables. Equations (4) and (5) no longer hold when measurements are correlated. The default settings in G*Power expects an f or η_p^2 that does *not* incorporate the correlation, while the correlation *is* incorporated in the output of software packages such as SPSS. One can enter the η_p^2 from SPSS output in G*Power after checking the "as in SPSS" checkbox in the options window, but forgetting this is a common mistake in power analyses for within designs in G*Power. For a one-way within-subject design, Cohen's f can be

converted into the Cohen's f SPSS uses through:

$$f^2_{SPSS} = f^2 \times \frac{k}{k-1} \times \frac{n}{n-1} \times \frac{1}{1-\rho} \tag{10}$$

and subsequently tranformed to η^2_p through Equation (5).

We can perform the simulation-based power analysis with the `ANOVA_power` or `ANOVA_exact` functions.

```
power_result_within_1 <- ANOVA_power(design_result_within_1,
                                     nsims = 100000)


exact_result_within_1 <- ANOVA_exact(design_result_within_1)
```

Revisiting our between-participant design, power was 81.14% when the enjoyment scores were uncorrelated. If we want to examine the power for a within design we need to enter a reasonable value for the correlation between dependent variables. Ideally this value is determined based on previous studies, and it often makes sense to explore a range of possible correlations in an a-priori power analysis. Let's assume our best estimate of the correlation between enjoyment ratings in a within-subject design is $r = 0.5$. The power for a repeated-measures ANOVA based on these values, where ratings for the three conditions are collected from 80 participants, is 98.38%. Because of the positive correlation between dependent variables, the effect size η^2_p is much larger for the within-subject design ($\eta^2_p = 0.12$) than for the 3 group between participants design ($\eta^2_p = 0.05$). Note that the Superpower package allow researchers to enter a correlation matrix that specifies the expected correlations between each individual pair of measurements, instead of assuming the correlations between all dependent variables are identical.

Power for Interactions

Let's assume the researcher plans to perform a follow-up experiment where in addition to making the voice sound cheerful or sad, a second factor is introduced by making the voice sound more robotic compared to the default human-like voice. Different patterns of results could be expected. Either the same effect is observed for robotic voices, or no effect is observed for robotic voices, or the opposite effect is observed for robotic voices (we enjoy a sad robotic voice more than a cheerful one, a "Marvin-the-Depressed-Robot Effect"). In the first case, we will only observe a main effect of voice, but in the other two scenarios there is an interaction effect between human-likeness of the voice and the emotional tone of the voice. We specify the design with the ANOVA_design function.

```
design_result_cross_80 <- ANOVA_design(
  design = "2b*2b",
  n = 80,
  mu = c(1, 0, 0, 1),
  sd = 2,
  labelnames = c("condition", "cheerful", "sad",
                 "voice", "human", "robot")
)
```

We can simulate a cross-over interaction for a 2x2 between-participant design with 80 participants in each group to examine the statistical power (see Figure 4 for the expected pattern of means). Again, we can do this with both simulation functions.

```
power_result_cross_80 <- ANOVA_power(design_result_cross_80,
                                     nsims = 100000)


exact_result_cross_80 <- ANOVA_exact(design_result_cross_80)
```
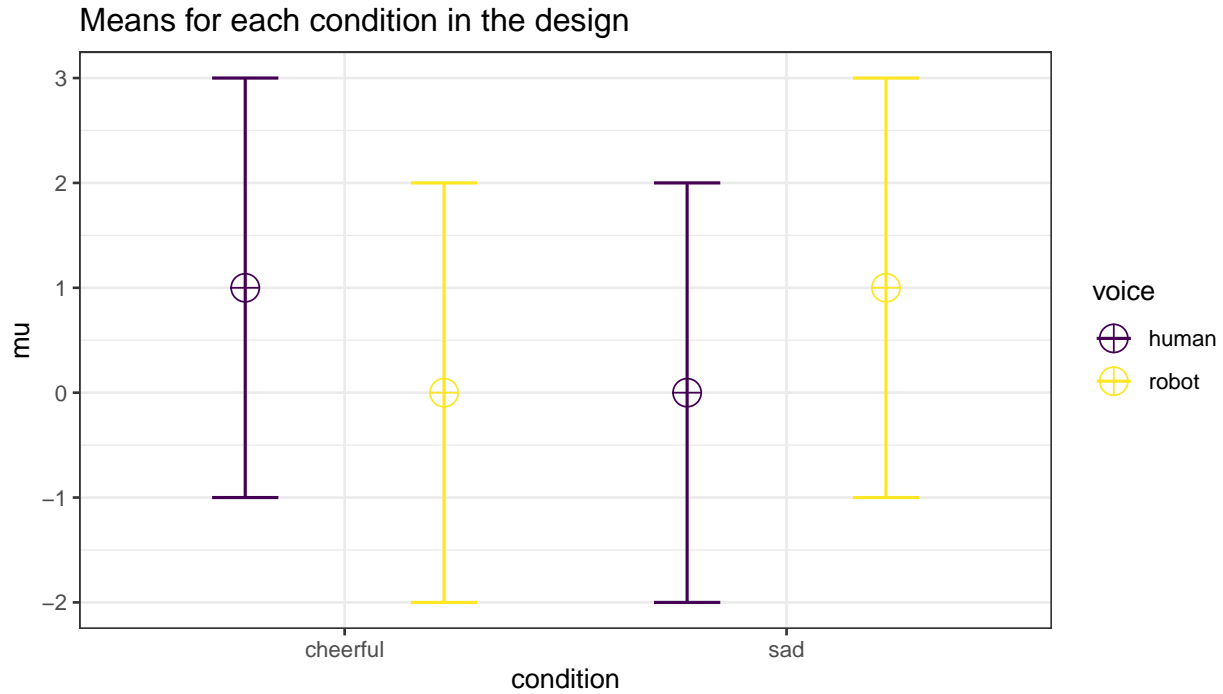
Figure 4. Vizualization for the expected means and standard deviations for a crossover interaction. Error bars represent one standard deviation.

Mathematically the interaction effect is computed as the cell mean minus the sum of the grand mean, the marginal mean in each row minus the grand mean, and the marginal mean in each column minus grand mean (see Maxwell, Delaney, and Kelley (2017)). For example, for the cheerful human-like voice condition this is 1 (the value in the cell) - (0.5 [the grand mean] + 0.5 [the cell mean minus the marginal mean in row 1] + 0.5 [the cell mean minus the marginal mean in column 2]). Thus, 1 - (0.5 + 0.5 + 0.5) = -0.5. Completing this for all four cells gives the values -0.5, 0.5, 0.5, -0.5. Cohen's f is then $f = \frac{\sqrt{\frac{-0.5^2 + 0.5^2 + 0.5 + -0.5^2}{4}}}{2} = 0.25$. Simulations show we have 99.38% power when we collect 80 participants per condition. Power is high, because we collected 80 participants in each condition. Compared to the two-group comparison with 80 participants per group a cross-over (also called "disordinal") interaction with two levels per factor has the same power as the initial two-group design if we halve the sample size per condition (i.e., n = 40 in each cell). This is accomplished simply by changing the `n` in the `ANOVA_design` function.

```r
design_result_cross_40 <- ANOVA_design(

  design = "2b*2b",

  n = 40,

  mu = c(1, 0, 0, 1),

  sd = 2,

  labelnames = c("condition", "cheerful", "sad",

                 "voice", "human", "robot"),

  plot = FALSE

)



And, power can be estimated with the simulation functions.
```

```r
power_result_cross_40 <- ANOVA_power(design_result_cross_40,

                                     nsims = 100000)



exact_result_cross_40 <- ANOVA_exact(design_result_cross_40)
```

Once we make that change, power with 40 participants per condition is 88.31%. Main effects in an ANOVA are based on the means for one factor averaged over the other factors (e.g., the main effect of a human-like versus robot-like voice, irrespective of whether it is cheerful or sad). The interaction effect can be contrast coded as 1, -1, -1, 1, and thus tests the scores of 80 participants against 80 other participants, just as for the *t*-test or the main effects. The key insight here is that not the sample size per condition, but the total sample size over all other factors determines the power for the main effects and the interaction (cf. J. Westfall, 2015b).

We can also examine the statistical power for a pattern of results that indicated that there was no difference in interacting with a cheerful of sad conversational agent with a robot

voice. In this case, we expect an "ordinal" interaction (the means for the human-like voice are never lower than the means for the robot-like voice, and thus there is no cross-over effect). The expected pattern of means is 1, 0, 0, 0, with only a single mean that differs from the others, and we create this new design.

```
design_result_ordinal <- ANOVA_design(
  design = "2b*2b",
  n = 160,
  mu = c(1, 0, 0, 0),
  sd = 2,
  labelnames = c("condition", "cheerful", "sad",
                 "voice", "human", "robot")
)
```

As has been pointed out (Giner-Sorolla, 2018; Simonsohn, 2014) these designs require larger samples sizes to have the same power to detect the interaction, compared to the two-group comparison. The reason for this is that the effect size is only half as large, with Cohen's f = 0.125 (compared to 0.25 in the cross-over interaction).

By steadily increasing the sample size in the simulation, we can examine the required sample size to get the same power as for the cross-over interaction with 40 participants per cell (88.31% power). A total sample size of 635 is required, almost four times as large as the total sample size for the two-group comparison (160). To make it easier to find the sample size per group that is required to calculate the expected power the plot_power function in Superpower can be used to plot the power for a specific design across a range of sample sizes (see Figure 6; code included below).

```
plot_data <- plot_power(design_result_ordinal,

                        min_n = 10, max_n = 200,

                        plot = TRUE)
```
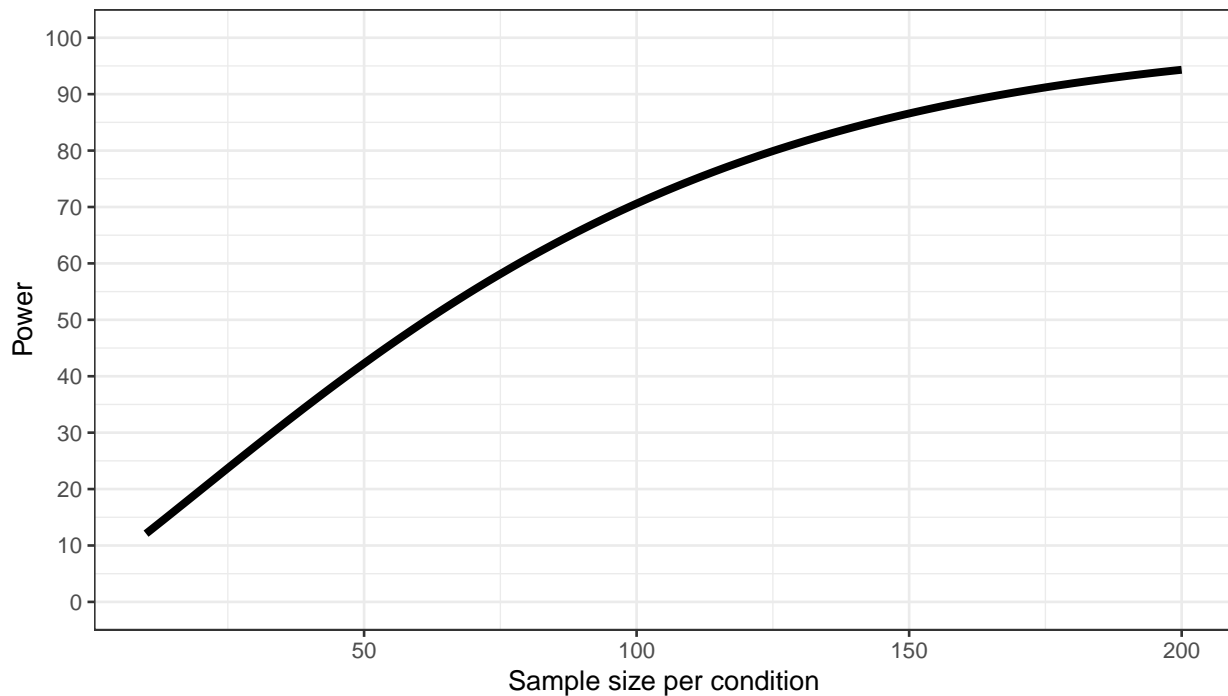

Figure 5. Power curves across a range of sample sizes per group from n = 10 to n = 200 for the expected main effects and ordinal interaction.

Where a two group comparison has a Cohen's f of 0.25 when the means are 1 and 0, a 2x2 ordinal interaction has a Cohen's f of 0.25 when the pattern of means is 2, 0, 0, 0, and a 2x2x2 interaction where only one cell differs from the other means has a Cohen's f of 0.25 when the pattern of means is 4, 0, 0, 0, 0, 0, 0, 0 across the eight cells. The take-home message is that a "medium" effect size translates into a much more extreme pattern of means in an ordinal interaction than in a disordinal (crossover) interaction, or in a 2x2x2 interaction compared to a 2x2 interaction (see also Perugini et al. (2018)). It might therefore be more intuitive to perform a power analysis based on the expected pattern of means, and compute Cohen's f based on this pattern, than to specify an effect size for the power analysis directly.

Power for Individual Comparisons

Although an initial goal when performing an ANOVA might be to test the *omnibus null hypothesis*, which answers the question whether there are *any* differences between group means, we often want to know which conditions differ from each other. Thus, an ANOVA is often followed up by individual comparisons (whether *planned* or *post-hoc*). One feature of Superpower is that it provides the statistical power for all individual comparisons that can be performed. The power and effect size estimates are based on simple t-tests. Taking into account the variance estimates from other groups through comparing the estimated marginal means (Lenth, 2019) (which the t-tests do not do) can have power benefits (Maxwell et al., 2017), and these estimated marginal means comparisons can be accomplished by setting `emm = TRUE` within the ANOVA_power or ANOVA_exact functions. However, these depend on whether the homogeneity assumption is met, which may not be warranted in psychological research, and violations of the homogeneity assumption impact Type 1 error rates (Delacre, Lakens, Mora, & Leys, 2018). The degree to which the violation affects Type 1 error rate can be estimated with the ANOVA_power function (see Assumptions section below)

Power analysis for individual comparisons is relatively straightforward and can easily be done in all power analysis software, but providing power for all individual comparisons alongside the ANOVA hopefully nudges researchers to take into account the power for follow-up tests. Sometimes the interaction in the ANOVA will have more power than the individual comparisons (i.e., in the case of a cross-over interaction) and sometimes the power for the interaction will be lower than the power for independent comparisons. It is always important to check if you have adequate power for all tests you plan to perform, which includes main effects, interactions, and individual comparisons performed as follow-up tests.

Type 1 Error Control in Exploratory ANOVA's

In a 2x2x2 design, an ANOVA will give the test results for three main effects, three two-way interactions, and one three-way interaction. Because seven statistical tests are performed, the probability of making at least one Type 1 error in a single exploratory 2x2x2 ANOVA is $1 - (0.95)^7 = 30\%$. It is therefore important to control error rates when performing multiple comparisons (Cramer et al., 2016).

It is possible to control the overall Type 1 error rate by specifying an adjustment for multiple comparisons in the simulation. By adjusting for multiple comparisons we ensure that we do not conclude there is an effect in *any* of the individual tests more often than the desired Type 1 error rate. Several techniques to control error rates exist, of which the best known is the Bonferroni adjustment. The Holm procedure is slightly more powerful than the Bonferroni adjustment, without requiring additional assumptions (for other approaches, see Bretz, Hothorn, & Westfall, 2011). Because the adjustment for multiple comparisons lowers the alpha level, it also lowers the statistical power. When designing a study where you will perform multiple comparisons, it is important to take into account the corrected alpha level when determining the required sample size, both for the ANOVA tests, as for the individual comparisons.

If we revisit the 2x2 between-condition ANOVA where a cross-over interaction was predicted and 40 participants per condition were collected, and apply a Holm correction for multiple comparisons.

```
design_result_holm_correction <-
  ANOVA_design(
    design = "2b*2b",
    n = 40,
    mu = c(1, 0, 0, 1),
    sd = 2,
```

```
    labelnames = c("condition", "cheerful", "sad",

                   "voice", "human", "robot")
 )
```

And we perform a Monte Carlo simulation to obtain the power analysis. Please note that the ANOVA_power function must used for situations where we want to simulate power when multiple comparisons corrections are applied.

```
power_result_holm_correction <- ANOVA_power(design_result_holm_correction,

                                            p_adjust = "holm",

                                            nsims = 100000)
```

We see power for the interaction is reduced from 88.31% without applying a correction, to 77.28% after correcting for multiple comparisons among the F-tests. There are three tests for a 2x2 design (two main effects, one interaction), but 6 individual comparisons between the four conditions, and thus the correction for multiple comparisons has a stronger effect on the individual comparisons. For example, the power for the t-test comparing two between-participant conditions falls from 60.17% without correcting for multiple comparisons, to 35.86% when adjusting the alpha level for the 6 possible individual comparisons between the four conditions. Please note, if a researcher would prefer to compare the estimated marginal means the R code could be simply modified (see below). The results of this code, saved under `power_result_holm_correction$emm_results`, would will be approximately the same as the individual comparisons above.

```
power_result_holm_correction <- ANOVA_power(design_result_holm_correction,

                                            p_adjust = "holm",

                                            emm = TRUE,

                                            emm_model = "univariate",
```

```
                                        contrast_type = "pairwise",

                                        emm_comp = "condition+voice",

                                        emm_p_adjust = "holm",

                                        nsims = 100000)
```

These simulations should reveal the cost (in terms of the statistical power) of exploring across all possible comparisons while controlling error rates. It is often advisable to preregister specific comparisons or tests that are of interest to a researcher. If one is really interested in all possible comparisons, but wants to control the probability of incorrectly acting as if there is an effect, the reduction in power due to the lower alpha level should be counteracted by increasing the sample size to maintain an adequate level of power for both the ANOVA as the individual comparisons.

Violation of Assumptions

So far in manuscript we have shown how simulations can be useful for ANOVA power analysis purposes. However, the utility of Monte Carlo simulations extends beyond simple power analysis or general sample size planning purposes. In Monte Carlo simulations we *can* generate data that, when analyzed via ANOVA, violate the assumptions of underlying an ANOVA. In fact, it is common to see Monte Carlo simulation studies published in the statistics literature when comparing the performance of various statistical analysis techniques in certain scenarios that may affect the type 1 error rate.

As an example, we could use Superpower to estimate the impact of violating the assumption of sphericity on the type 1 error rate for a simple one-way repeated measures ANOVA. For most experiments in real life, this assumption could be violated since the correlations between levels within factors may vary. This is problematic because the univariate ANOVAs assume sphericity, which means, for all intents and purposes, that the correlations between factor-levels are equal and the standard deviations at each level are

equal as well. This assumption is tenuous at best and is typically "adjusted" for by applying a sphericity correction (e.g, Greenhouse-Geisser or Huynh-Feldt adjusted output).

An inquiring researcher may wonder, how much will this violation inflate the type 1 error rate? We can simulate an example (below) wherein there are no differences between the repeated measures and varying corrlations between the levels (r = .05-.90) of repeated measures. Again, we set this up in the ANOVA_design function.

```
design_result_s1 <- ANOVA_design(
  "4w",
  n = 29,
  r = c(.05, .15, .25, .55, .65, .9),
  sd = 1,
  mu = c(0, 0, 0, 0)
)
```

We then perform the Monte Carlo simulation using the ANOVA_power function.

```
power_result_s1 <- ANOVA_power(design_result_s1, nsims = 100000)
```

The simulated type 1 error rate for the univariate ANOVA is 7.33%. The multivariate ANOVA (MANOVA) does not assume sphericity and therefore should be robust to this pattern of correlations. We can then check the MANOVA results and see this analysis approach keeps the type 1 error rate at 4.86%. Alternatively, we could re-run the simulation with a Greenhouse-Geisser (`correction = "GG"`) or Huynh-Feldt (`correction = "HF"`) corrections for sphericity the typical univariate ANOVA to determine how well these procedures would preserve the type 1 error rate as well. After we determine which analysis approach best preserves the type 1 error rate we could then re-run the simulation with the pattern of means we want to detect and estimate the power for the given design.

P-value distributions

Statistical power is the long run probability of observing a *p*-value smaller than the alpha level. One intuitive way to illustrate this concept is to plot the distribution of *p*-values for all simulations. The ANOVA_power function automatically stores plots for *p*-value distributions for each test in the simulation. In Figure 6 we see that for the initial 2 group between-participant design most *p*-values fall below the alpha level (in the figure, the vertical line as 0.05). We can produce by calling the stored results from the ANOVA_power function and selecting the plot1 object.

```
power_result$plot1
```

For the 1e+05 simulations, 88.19% fall below the alpha level. As power is increased the p-value distribution becomes steeper, and when there is no effect the *p*-values are uniformly distributed.
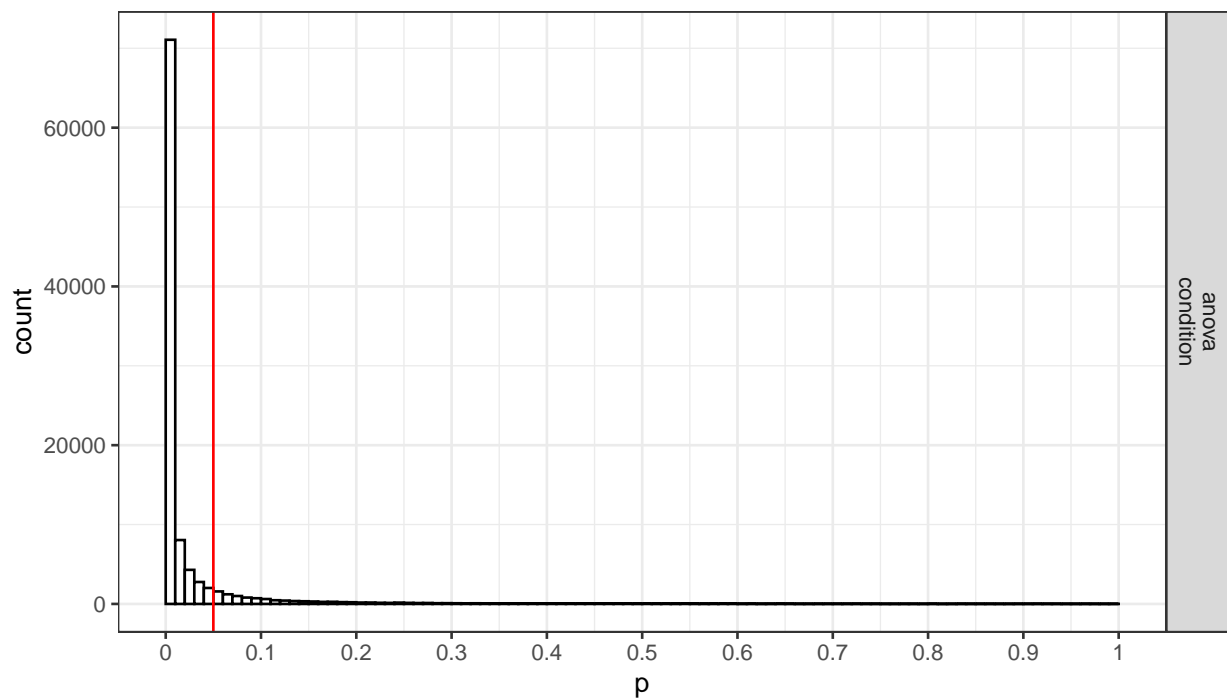
Figure 6. Distribution of p-values for the initial two-group between subjects design.

Conclusion

It is important to carefully justify the sample size when designing informative studies. Simulation based approaches can help to provide insights into the factors that affect the statistical power for factorial ANOVA designs. The effect of violating the assumptions of ANOVAs can also be evaluated using a simulation based approach. Exploring the power for designs with specific patterns of means, standard deviations, and correlations between variables can be used to choose a design and sample size that provides the highest statistical power for future studies. The R package (https://github.com/arcaldwell49/Superpower), guide book (https://arcaldwell49.github.io/SuperpowerBook), and Shiny app (http://shiny.ieis.tue.nl/anova_power/) that accompany this paper enable researchers to perform simulations for factorial experiments of up to three factors and any number of levels, making it easy to perform simulation-based power analysis without extensive programming experience.

Author Contributions

D. Lakens and A. R. Caldwell collaboratively developed the Superpower R package. D. Lakens wrote the initial draft, and both authors revised the manuscript. A. R. Caldwell created the Shiny apps.

ORCID iD's

Daniël Lakens ⬤ https://orcid.org/0000-0002-0247-239X Aaron R. Caldwell ⬤ https://orcid.org/0000-0002-4541-6283

Acknowledgements

Chris Aberson's pwr2ppl package. We are grateful to Jonathon Love for the development of a jamovi module based on Superpower which should be released soon.

Declaration of Conflicting Interests

Funding

Open Practices

The code to reproduce the analyses reported in this article has been made publicly available via the Open Science Framework and can be accessed at https://osf.io/pn8mc/.

References

Aberson, C. L. (2019). *Applied Power Analysis for the Behavioral Sciences* (2nd ed.). New York: Routledge.

Bretz, F., Hothorn, T., & Westfall, P. H. (2011). *Multiple comparisons using R*. Boca Raton, FL: CRC Press.

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with some simple guidelines. *Journal of Cognition.*

Campbell, J., & Thompson, V. A. (2012). MorePower 6.0 for anova with relational confidence intervals and bayesian analysis. *Behavior Research Methods*, *44*, 1255–1265. Retrieved from https://doi.org/10.3758/s13428-012-0186-0

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, N.J: L. Erlbaum Associates.

Cramer, A. O. J., van Ravenzwaaij, D., Matzke, D., Steingroever, H., Wetzels, R., Grasman, R. P. P. P., . . . Wagenmakers, E.-J. (2016). Hidden multiplicity in exploratory multiway ANOVA: Prevalence and remedies. *Psychonomic Bulletin & Review*, *23*(2), 640–647. doi:10.3758/s13423-015-0913-5

Delacre, M., Lakens, D., Mora, Y., & Leys, C. (2018). Why Psychologists Should Always Report the W-test Instead of the F-Test ANOVA. *PsyArXiv.* doi:10.17605/OSF.IO/WNEZG

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). GPower 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. doi:10.3758/BF03193146

Giner-Sorolla, R. (2018, January). Powering Your Interaction. *Approaching Significance.* https://approachingblog.wordpress.com/2018/01/24/powering-your-interaction-2/.

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269. doi:10.1177/2515245918770963

Lang, A.-G. (2017). Is intermediately inspecting statistical data necessarily a bad research practice? *The Quantitative Methods for Psychology.*

Lenth, R. (2019). Emmeans: Estimated marginal means, aka least-squares means. Retrieved from https://CRAN.R-project.org/package=emmeans

Maxwell, S. E., Delaney, H. D., & Kelley, K. (2017). *Designing Experiments and Analyzing Data: A Model Comparison Perspective, Third Edition* (3rd ed.). New York, NY: Routledge.

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample Size Planning for Statistical Power and Accuracy in Parameter Estimation. *Annual Review of Psychology, 59*(1), 537–563. doi:10.1146/annurev.psych.59.103006.093735

Perugini, M., Gallucci, M., & Costantini, G. (2018). A Practical Primer To Power Analysis for Simple Experimental Designs. *International Review of Social Psychology, 31*(1), 20. doi:10.5334/irsp.181

Simonsohn, U. (2014, March). No-way Interactions. *Data Colada.* http://datacolada.org/17.

Westfall, J. (2015a). PANGEA: Power analysis for general anova designs. *Unpublished Manuscript. Available at Http://Jakewestfall. Org/Publications/Pangea. Pdf.*

Westfall, J. (2015b, May). Think about total N, not n per cell. *Cookie Scientist.* http://jakewestfall.org/blog/index.php/2015/05/26/think-about-total-n-not-n-per-cell/.