Special Section: Using Simulation to Convey Statistical Concepts
Tutorial

ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Advances in Methods and Practices in Psychological Science
January-March 2021, Vol. 4, No. 1, pp. 1–14
© The Author(s) 2021
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2515245920951503
www.psychologicalscience.org/AMPPS
SAGE

Simulation-Based Power Analysis for Factorial Analysis of Variance Designs

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

[1]Human-Technology Interaction Group, Eindhoven University of Technology; [2]Department of Health, Human Performance and Recreation, University of Arkansas; and [3]Thermal and Mountain Medicine Division, U.S. Army Research Institute of Environmental Medicine, Natick, Massachusetts

Abstract

Researchers often rely on analysis of variance (ANOVA) when they report results of experiments. To ensure that a study is adequately powered to yield informative results with 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 allow power analyses for complex designs with several within-participants factors. Moreover, power analyses often need η_p^2 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 online Shiny apps to enable researchers without extensive programming experience to perform simulation-based power analysis for ANOVA designs of up to three within- or between-participants factors. Predicted effects are entered by specifying means, standard deviations, and, for within-participants factors, the correlations. The simulation provides the statistical power for all ANOVA main effects, interactions, and individual comparisons. The software can plot power across a range of sample sizes, can control for multiple comparisons, and can compute power when the homogeneity or sphericity assumption is violated. This Tutorial demonstrates how to perform a priori power analysis to design informative studies for main effects, interactions, and individual comparisons and highlights important factors that determine the statistical power for factorial ANOVA designs.

Keywords

power analysis, ANOVA, hypothesis test, sample-size justification, open materials

Received 5/28/19; Revision accepted 5/14/20

When a researcher aims to test hypotheses with an analysis of variance (ANOVA), the sample size of the study should be justified on the basis of 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 the statistical power of a test is low, there is a high probability of a Type II error, or concluding there is no effect when a true effect exists in the population of interest.

Several excellent resources that explain power analyses are available. These include books (Aberson, 2019; Cohen, 1988), general reviews (Maxwell et al., 2008), and practical primers (Brysbaert, 2019; Perugini et al., 2018). Whereas power analyses for individual comparisons are relatively easy to perform, power analyses for factorial ANOVA designs are a bigger challenge. There is a range of power-analysis software available, such as G*Power (Faul et al., 2007), MorePower (Campbell & Thompson, 2012), PANGEA (Westfall, 2015a), *pwr2ppl* (Aberson, 2019), APRIOT (Lang, 2017), PASS (NCSS LLC, Kaysville, UT), and SAS (SAS Institute, Cary, NC). These tools differ in their focus (e.g., sequential analyses for APRIOT, linear mixed models for PANGEA), the tests they provide power analyses for (e.g., whether they allow violations of the homogeneity assumption or unequal sample sizes, whether they can be used with analysis of covariance [ANCOVA] designs), and the input they require (e.g., effect sizes, raw data, or means, standard deviations, correlations, and sample sizes).[1] Despite

Corresponding Author:
Daniël Lakens, Human-Technology Interaction Group, Eindhoven University of Technology
E-mail: D.Lakens@tue.nl

this wide range of software options, in our experience researchers often struggle to perform power analyses for ANOVA designs.

In this article, we introduce the *Superpower* R package and accompanying Shiny apps, which use simulations to perform power analyses for factorial ANOVA designs. We designed *Superpower* with the goal for it to be free, to be available both as R functions and as an online app, and to easily allow researchers to perform power analyses for a wide range of ANOVA designs. Compared to G*Power, the *pwr* R package (Champely, 2020), and the *pwr2ppl* R package, *Superpower* can compute power for a wider range of designs (e.g., up to three factors with 999 levels). Compared to PANGEA, G*Power, and MorePower, *Superpower* requires input that we believe is somewhat more intuitive, as users enter means, standard deviations, and correlations, instead of effect sizes and variance components. A unique feature of *Superpower* is that it allows users to easily correct for multiple comparisons in exploratory ANOVA designs, and that it automatically provides the statistical power for all main effects, interactions, and simple comparisons for a specified ANOVA design. The online manual at http://arcaldwell49.github.io/SuperpowerBook (Caldwell et al., 2020) provides detailed examples of power analyses for a variety of designs (ranging from one-way ANOVA designs to three-way interactions in mixed designs, multivariate analyses of variance [MANOVAs], and situations in which ANOVA assumptions are violated), as well as examples validating power analyses in *Superpower* against existing software. A current limitation of *Superpower* is that it cannot compute power for ANCOVAs or linear mixed models.

Superpower allows researchers to perform simulation-based power analyses without having extensive programming knowledge. 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 focusing on the *F* test, we illustrate through simulations how the power of factorial ANOVA designs depends on the pattern of means across conditions, the number of factors and levels, the sample size, and whether the alpha level needs to be controlled for multiple comparisons.

Disclosures

The code to reproduce the analyses reported in this article has been made publicly available via OSF and can be accessed at https://osf.io/pn8mc/. An online manual for *Superpower* can be accessed at https://aaroncaldwell.us/SuperpowerBook/. In addition, there are shiny apps for the `ANOVA_exact` (https://arcstats.io/shiny/anova-exact/) and `ANOVA_power` (https://arcstats.io/shiny/anova-power/) functions mentioned throughout this article. The *Superpower* R package is available on CRAN (https://CRAN.R-project.org/package=Superpower), and experimental versions of the package are available on our GitHub repository (https://github.com/arcaldwell49/Superpower).

A Basic Example

Imagine that we perform a study in which participants interact with an artificial voice assistant who sounds either cheerful or sad. We measure how much 80 participants in each condition enjoy interacting with the voice assistant by collecting responses on a line scale (coded continuously from −5 to 5). We observe a mean of 0 in the sad condition and a mean of 1 in the cheerful condition, and the estimated standard deviation is 2. After we submit the manuscript for publication, reviewers ask us to add a study with a neutral control condition to examine whether cheerful voices increase enjoyment or sad voices decrease enjoyment (or both). Depending on what the mean enjoyment in the neutral condition in the population is, what sample size would we need for a high-powered test of the expected pattern of means? A collaborator suggests switching from a between-participants design to a within-participants design to collect data more efficiently. What impact will switching to a within-participants design have on the required sample size? The effect size observed in the first study is sometimes referred to as a “medium” effect size, according to the benchmarks by Cohen (1988). Does it make sense to perform an a priori power analysis for a medium effect size if we add a third between-participants condition or switch to a within-participants ANOVA design? And if we justify the sample size on the basis of 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)? Before we answer these questions, let us review some of the basic concepts of statistical power and examine how power calculations are typically performed.

Calculating Power for ANOVA Designs

Let us consider the two-condition design described earlier, in which enjoyment is measured among 80 participants per condition who 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. To perform an a priori power analysis, researchers need to specify an effect size for the alternative hypothesis (for details on effect-size calculations, see Box 1). Figure 1 and Figure 2 show the distributions of the effect sizes—Cohen’s *d*

Box 1. Formulas for Effect Sizes for Analysis of Variance Designs

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

$$d = \frac{\bar{x}_1 - \bar{x}_2}{\sigma_p}. \tag{1}$$

The generalization of Cohen's d to more than two groups is Cohen's f, which is the standard deviation of the means divided by the standard deviation (Cohen, 1988), or

$$f = \frac{\sigma_{\bar{x}}}{\sigma}, \tag{2}$$

where for equal sample sizes,

$$\sigma_m = \sqrt{\frac{\sum_{i=1}^{k}\left(\bar{x}_i - \bar{x}\right)^2}{k}}. \tag{3}$$

For two groups, Cohen's f is half as large as Cohen's d, or $f = ½d$. In power-analysis software, the input is often η_p^2, which can be converted into Cohen's f:

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

Likewise, Cohen's f can be converted into η_p^2:

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

Power calculations rely on the noncentrality parameter (λ). In a between-participants one-way analysis of variance, λ is calculated as

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

where f is Cohen's f and N is the total sample size.

for the t test and η_p^2 for the F test—that should be observed when there is no effect and when the alternative hypothesis is true (d = 0.5 and η_p^2 = .0588, respectively).[2] In each figure, the light-gray areas under the distribution for the null hypothesis mark the observed effect sizes that would lead to a Type I error (observing a statistically significant result if the null hypothesis is true), and the dark-gray area under the curve for the distribution under the alternative hypothesis marks the observed effect sizes that would lead to a Type II error (observing a nonsignificant result when there is a true effect).

A test result is statistically significant when the p value is smaller than the alpha level or when the test statistic (e.g., an F value) is larger than a critical value. For a given sample size, we can also calculate a critical *effect size*, and a result is statistically significant if the observed effect size is more extreme than the critical effect size. Given the sample size of 80 participants per group, observed effects are statistically significant when $\hat{d}$ is larger than 0.31 in a t test or $\hat{\eta}_p^2$ is larger than .024 in an F test (see the vertical dashed lines in Fig. 1 and Fig. 2). The goal of an a priori power analysis is to determine the sample size required, in the long run, to observe a p value smaller than the chosen alpha level with a predetermined probability, given an assumption about the true population effect size. To calculate the sample size

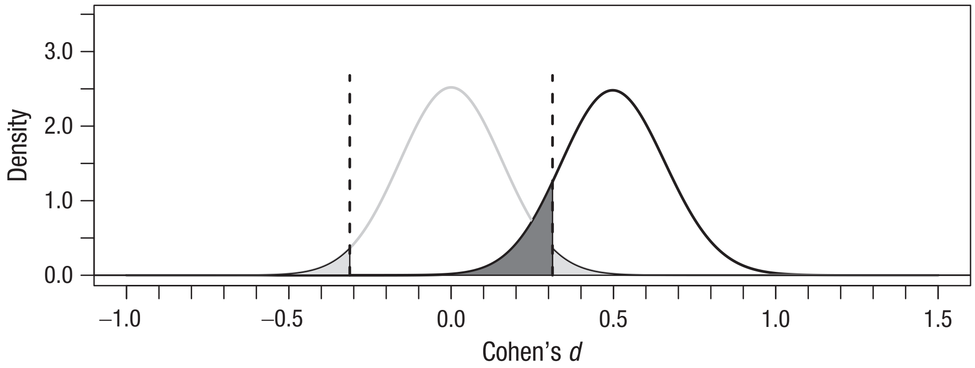


Fig. 1. Distribution of Cohen's *d* under the null hypothesis (gray curve) and the alternative hypothesis assuming d = 0.5 in the population (black curve), given n = 80 per condition. The shaded areas indicate the observed effect sizes that would lead to Type I (light gray) and Type II (dark gray) errors, and the vertical lines indicate the critical effect sizes. See the text for additional explanation.

required to reach a desired statistical power, one has to specify the alternative hypothesis and the alpha level. Given λ (the noncentrality parameter, which together with the degrees of freedom specifies the shape of the expected effect-size distribution under a specified alternative hypothesis, illustrated by the black curves in Figs. 1 and 2), we can calculate the area under the curve that is more extreme than the critical effect size (i.e., in Fig. 2, the area to the right of the critical effect size). Under the alternative hypothesis that the true population effect size is 0.5 (*d*) or .0588 (η_p^2), if data are collected from 80 participants in each condition, and an alpha of .05 is used, in the long run 88.16% of the tests will yield an effect size that is larger than the critical effect size.

Power Calculations in *Superpower*

Superpower can be used in R (run `install.packages("Superpower")`) or in the online Shiny apps (see https://arcstats.io/shiny/anova-exact/ and https://arcstats.io/shiny/anova-power/). The code underlying the *Superpower* R package and the Shiny apps generates data for each condition in the design and performs an ANOVA and *t* tests for all comparisons between conditions. The simulation can be based on any design specified using the `ANOVA_design` function, the result of which is stored and passed on to either of the two functions to compute power. Users specify the design by indicating the number of levels for each factor (e.g., `2`) and whether the factor is manipulated within (`w`) or between (`b`) participants. *Superpower* can handle up to three factors (separated by `*`). A `2b` design means a single factor with two groups is manipulated between participants, whereas a `2b*2w` design is a 2 × 2 mixed ANOVA in which the first factor is manipulated between and the second within participants. Users also specify the sample size per condition (`n`), the predicted pattern of means across all conditions, the expected standard deviation, and the correlation between variables (for within-participants designs). To make it easier

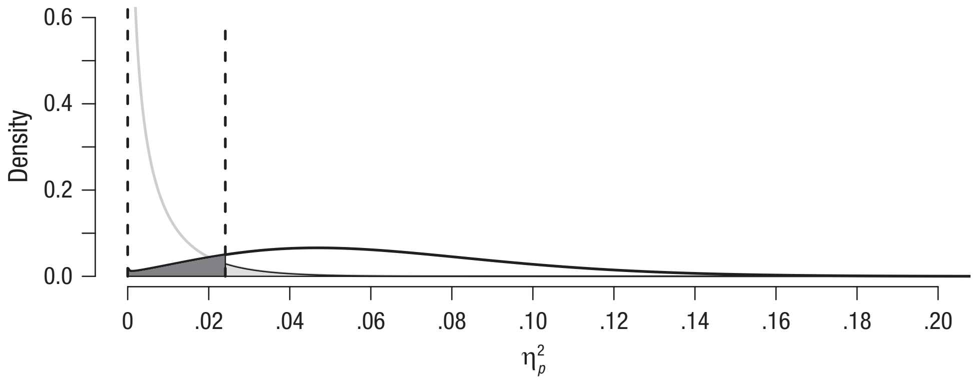


Fig. 2. Distribution of η_p^2 under the null hypothesis (gray curve) and the alternative hypothesis assuming η_p^2 = .0588 in the population (black curve), given n = 80 per condition. The shaded areas indicate the observed effect sizes that would lead to Type I (light gray) and Type II (dark gray) errors, and the vertical lines indicate the critical effect sizes. See the text for additional explanation.

to interpret the output, users can specify factor names and names for each factor's levels (e.g., condition, cheerful, sad).

An example of the R code is

```
design_result <- ANOVA_design(
  design = "2b", n = 80,
  mu = c(1, 0), sd = 2,
  labelnames = c("condition",
                 "cheerful", "sad"),
  plot = TRUE)
```

For a visual confirmation of the input, the R function creates a figure that displays the means and standard deviation (see the right side of Fig. 3). After the design has been specified, there are two ways to calculate the statistical power of an ANOVA through simulations. The ANOVA_power function simulates data sets repeatedly according to the specified parameters and calculates the percentage of statistically significant results. Following is the code for performing 1,000 simulations, which should take approximately 15 s and yields reasonably accurate results for experimenting with the power-analysis function:

```
result_monte <- ANOVA_power(design_result,
                            nsims = 1000)
```

For most designs, increasing the number of simulations to 10,000, which means the calculations would take a few minutes to complete, should give results accurate enough for most practical purposes.

The ANOVA_exact function simulates a data set that has *exactly* the desired properties, performs an ANOVA, and uses the ANOVA results to compute the statistical power.

Here is an example of the same design stated above:

```
result_exact <- ANOVA_exact(design_result)
```

The first approach is a bit more flexible (e.g., it allows for sequential corrections for multiple comparisons, such as the Holm procedure), but the second approach is much faster (and generally recommended). There is often uncertainty about the values that are required to perform an a priori power analysis. The true (population-level) pattern of means, standard deviations, and correlations is unknown (and the goal of the study is to learn what this data pattern looks like). It makes sense to examine power across a range of assumptions, from more optimistic scenarios, to more conservative estimates. In many cases, researchers should consider using a sample size that guarantees sufficient power for the smallest effect size of interest, instead of the effect size they expect. (For examples of ways to specify a smallest effect sizes of interest, see Lakens et al., 2018). This approach ensures that the study can be informative, even when there is uncertainty about the true effect size.

If ANOVA_power is used, the results from the simulation will vary each time the simulation is performed (unless a seed is specified, e.g., set.seed = 2019). A user should specify the number of simulations (the more simulations, the more accurate the results are, but the longer the simulation takes), the alpha level for the tests, and any adjustments for multiple comparisons that are required. The outputs from ANOVA_exact and ANOVA_power are similar, and provide the statistical power for the ANOVA and all simple comparisons between conditions. Here is an example of the output from ANOVA_power:

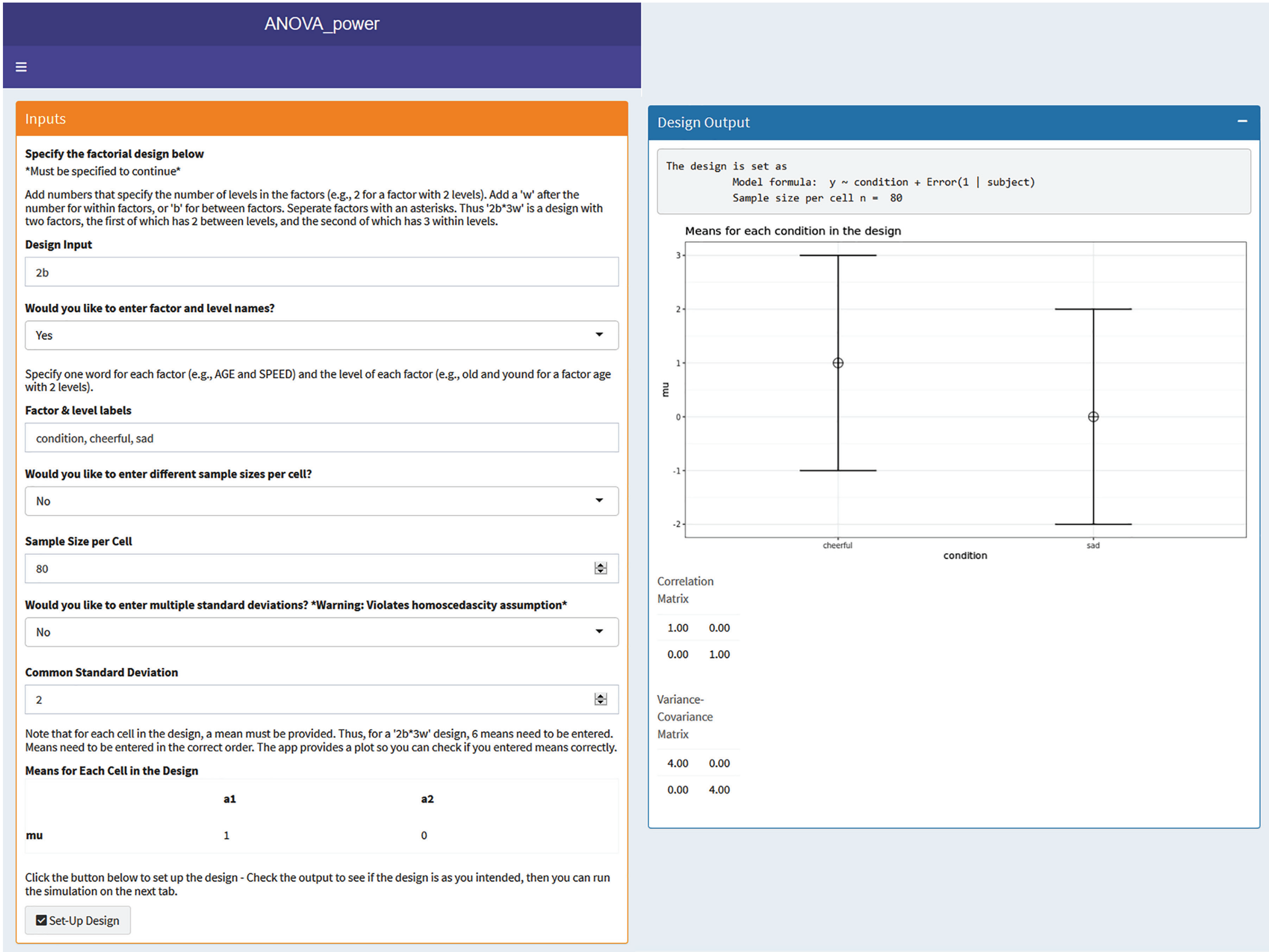
```
Power and Effect sizes for ANOVA tests
                   power  effect_size
anova_condition   88.191   0.06425
Power and Effect sizes for
pairwise comparisons (t-tests)
                  power  effect_size
p_cheerful_sad  88.191   -0.5017
```

The same results are returned in the online Shiny app, but users can also choose a "download PDF report" option to receive the results as a PDF file that can be saved to be included as documentation for sample-size requirements (e.g., for a preregistration, Registered Report, or grant application). An example of the input in the ANOVA_power Shiny app and the corresponding results are presented in Figures 3 and 4.

These results show that when 100,000 simulations are performed for our two-group between-participants design with means of 1 and 0, a standard deviation of 2, and 80 participants in each group (for a total of 160 participants), with a seed set to 2019 (these settings were used for all simulation results reported in this article), the statistical power (based on the percentage of $p < \alpha$ results) is 88.19% and the average $\hat{\eta}_p^2$ is .064. The simulation also provides the *t*-test results for the individual comparisons. Since there are only two groups in this example, the statistical power for the individual comparisons is identical to that for the ANOVA, but the expected effect size is given as Cohen's $\hat{d}$: −0.50.

Simulating Statistical Power for Different Factorial Designs

Now that we have illustrated the basic idea behind power analyses in *Superpower*, we can use it 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 first examine what happens if we



Specify the factorial design below
Must be specified to continue

Add numbers that specify the number of levels in the factors (e.g., 2 for a factor with 2 levels). Add a 'w' after the number for within factors, or 'b' for between factors. Seperate factors with an asterisks. Thus '2b*3w' is a design with two factors, the first of which has 2 between levels, and the second of which has 3 within levels.

Design Input

2b

Would you like to enter factor and level names?

Yes ▾

Specify one word for each factor (e.g., AGE and SPEED) and the level of each factor (e.g., old and yound for a factor age with 2 levels).

Factor & level labels

condition, cheerful, sad

Would you like to enter different sample sizes per cell?

No ▾

Sample Size per Cell

80

Would you like to enter multiple standard deviations? *Warning: Violates homoscedascity assumption*

No ▾

Common Standard Deviation

2

Note that for each cell in the design, a mean must be provided. Thus, for a '2b*3w' design, 6 means need to be entered. Means need to be entered in the correct order. The app provides a plot so you can check if you entered means correctly.

Means for Each Cell in the Design

	a1	a2
mu	1	0

Click the button below to set up the design - Check the output to see if the design is as you intended, then you can run the simulation on the next tab.

```
The design is set as
          Model formula:  y ~ condition + Error(1 | subject)
          Sample size per cell n =  80
```

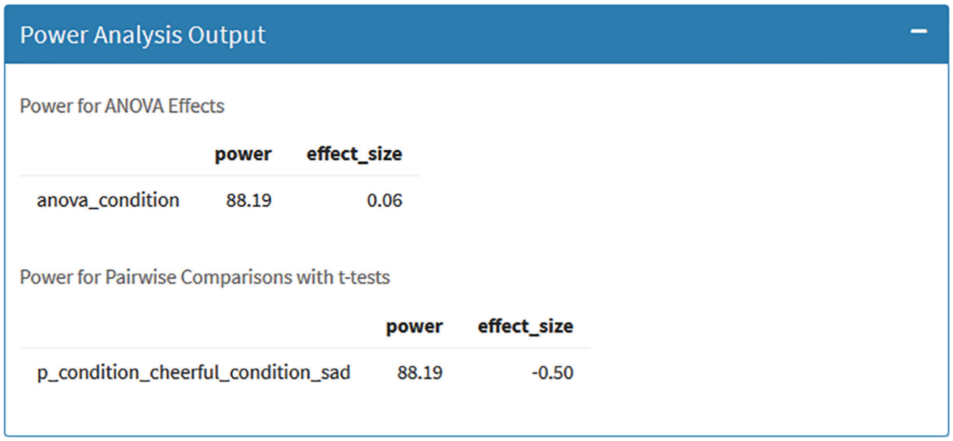
Correlation Matrix

1.00	0.00
0.00	1.00

Variance-Covariance Matrix

4.00	0.00
0.00	4.00

Fig. 3. Screenshots of the *ANOVA_power* Shiny app, illustrating inputs and the corresponding design output.



Power for ANOVA Effects

	power	effect_size
anova_condition	88.19	0.06

Power for Pairwise Comparisons with t-tests

	power	effect_size
p_condition_cheerful_condition_sad	88.19	-0.50

Fig. 4. Screenshot of the *ANOVA_power* Shiny app showing the results of the power analysis for the inputs in Figure 3.

add a third, neutral condition to the design. Let us assume that we expect the mean enjoyment rating for the neutral condition to fall either perfectly between the means in the cheerful and sad conditions or to be equal to the mean in the cheerful condition. Will simply collecting data from 80 additional participants in the neutral condition (for a total of 240 participants) be enough for a one-way ANOVA to have sufficient power? The R code to specify the first design is

```
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 three between-participants conditions, and we can explore what would happen if we collect data from 80 participants in each condition.

If we assume that the mean in the neutral condition falls exactly between the means in the cheerful and sad conditions, the simulations show that the statistical power for a three-group one-way ANOVA *F* test is reduced to 81.14%. If we assume that the mean in the neutral condition is equal to the mean in the cheerful condition, the power increases to 91.03%. This highlights how different expected patterns of means translate into different effect sizes, and thus different levels of statistical power. Compared to the two-group design (for which the power was 88.19%), three things have changed in the three-group design. First, the numerator degrees of freedom has increased because an additional group has been added to the design, which makes the noncentral *F* distribution more similar to the central *F* distribution, which reduces the statistical power. Second, the total sample size is 50% larger after 80 participants have been added 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 (if we expect the mean in the neutral condition to fall between the means in the other two conditions) or 0.24 (if we expect the mean in the neutral condition to equal the mean in the sad condition), which reduces the statistical power. The most important take-home message is that changing an experimental design can have several opposing effects on the power of a study, depending of the pattern of means. The exact effect of these three changes on the statistical power is difficult to anticipate from one design to the next. This highlights the importance of thinking about the specific pattern of means across conditions that a theory predicts when performing an a priori power analysis.

Power for Individual Comparisons

Although an initial goal might be to test the *omnibus null hypothesis* (i.e., ANOVA), which answers the question whether there are *any* differences among group means, researchers often want to know which specific conditions differ from each other. Thus, an ANOVA is often followed up by individual comparisons (whether planned or post hoc). It is very important that researchers consider whether their design will have enough power for any individual comparisons they want to make. *Superpower* automatically provides the statistical power for all individual comparisons that can be performed, so that researchers can easily check if their design is well powered for follow-up tests. By default, the power and effect-size estimates are based on simple *t* tests. In our hypothetical example, with expected means of 0, 0.5, and 1 for the cheerful, neutral, and sad conditions, statistical power is highest for the comparison between the cheerful and sad conditions (88.22%).

We see that (except for minor differences due to the fact that simulations will give slightly different results each time they are run) the power estimate is identical to that for the two-group design. The estimated statistical power provided by the `ANOVA_power` function is only 35.03% for the comparison of the cheerful and neutral conditions and 34.72% for the comparison of the sad and neutral conditions (the two power estimates differ slightly because they are based on simulations, even though the difference between means is identical, i.e., 0.5). It is clear that our design, despite having sufficient power to detect a main effect, is not well powered for the individual comparisons we are interested in.

It is also possible to combine variance estimates from all conditions and calculate the estimated marginal means (Lenth, 2019) when performing individual comparisons. This is done by setting `emm = TRUE` within the `ANOVA_power` or `ANOVA_exact` function, or checking this option in the Shiny app. This approach often has greater statistical power (Maxwell et al., 2017), depending on whether the assumption of equal variances (also known as the homogeneity assumption) is met, which may not be warranted in psychological research (Delacre et al., 2017). The degree to which violations of the homogeneity assumption affect Type I error rates can be estimated with the `ANOVA_power` function (see the Violation of Assumptions section). Power analysis for individual comparisons is relatively straightforward and can easily be done in all power-analysis software, but we hope that by providing power for all individual comparisons alongside the ANOVA result by default, *Superpower* and the Shiny apps will nudge researchers to take into account the power for follow-up tests.

When performing multiple individual comparisons, researchers need to choose the alpha level and ensure that the Type I error rate is not inflated. By adjusting for multiple comparisons, they ensure that they do not conclude there is an effect in any of the individual tests more often than the desired Type I error rate. Of the several techniques to control error rates, the best known is the Bonferroni correction. The Holm procedure is slightly more powerful than the Bonferroni correction, without requiring additional assumptions (for other approaches, see Bretz et al., 2011). Power analyses using a manually calculated Bonferroni correction can be performed with the `ANOVA_exact` function by specifying the adjusted alpha level, but the sequential Holm approach can be performed only in the Monte Carlo simulation approach (e.g., `ANOVA_power`). Because the adjustment for multiple comparisons lowers the alpha level, it also lowers statistical power. If we repeat the hypothetical ANOVA with three conditions while applying the Holm correction, we would have approximately 78% power for the expected difference between the cheerful and sad conditions after controlling for multiple comparisons with the Holm procedure (compared to 88.22% power without correcting for multiple comparisons), and only 26% power when we compare the cheerful and sad conditions with the neutral condition. As the number of possible paired comparisons increases, the alpha level is reduced, and power is reduced, all else being equal.

These power analyses reveal the cost (in terms of the statistical power) of exploring all possible paired comparisons while controlling error rates. To maintain an adequate level of power after lowering the alpha level to control the Type I error rate after multiple comparisons, the sample size should be increased. In a one-way ANOVA, multiple comparisons are an issue only for the follow-up comparison, but in a 2 × 2 × 2 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 I error in a single exploratory 2 × 2 × 2 ANOVA is $1 - (.95)^7$, or 30%. It is therefore important to control error rates in exploratory ANOVAs (Cramer et al., 2016). If a researcher is interested only in specific tests, it is advisable to preregister and test only these comparisons instead of correcting the alpha level for all possible comparisons (Haans, 2018).

Power for Within-Participants Designs

What would happen if we performed the second study as a within-participants design? Instead of collecting data from three groups of participants, we might collect data from only one group and let this group evaluate the cheerful, neutral, and sad voice assistants. If we want to examine the power for a within-participants design, we need to enter our best estimate for the true population value of the correlation between dependent measurements. Ideally this value is based on previous studies, and when there is substantial uncertainty about the true population value, it often makes sense to explore a range of plausible correlations. Let us assume that our best estimate of the correlation between enjoyment ratings in a within-participants design (ρ) is .5. The following `ANOVA_design` function specifies this design:

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

Note that the design specification has changed from `3b` (a one-factor between-participants design with three

Box 2. Formula for Effect Sizes for Within-Participants Designs

The effect size in a two-group within-participants design is referred to as Cohen's d_z (because it is the effect size of the difference score between x and y, referred to as z). The relation between the standard deviation of the mean difference and the standard deviation of the means is

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

Cohen's d_z is used in power analyses for dependent-samples t tests, but there is no equivalent Cohen's f_z for a within-participants analysis of variance, and Cohen's f is identical for within- and between-participants designs. Instead, the value for λ is adjusted based on the correlation. For a one-way within-participants design, λ is identical to the calculation for a between-participants design in Equation 6, multiplied by u, a correction for within-participants designs that is calculated as

$$u = \frac{k}{1-\rho}, \tag{9}$$

where k is the number of levels of the within-participants factor and ρ is the correlation between dependent variables. Equations 4 and 5 no longer hold when measurements are correlated. G*Power (Faul et al., 2007) by default expects the user to input an f or η_p^2 that does not incorporate the correlation, but 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-participants designs in G*Power. For a one-way within-participants design, Cohen's f can be converted into the Cohen's f SPSS uses through

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

and subsequently transformed to η_p^2 through Equation 5.

levels) to `3w` (a one-factor within-participants design with three levels), and the correlation parameter `r = 0.5` has been added to specify the expected correlation between dependent variables in the population.

A rough but useful approximation of the sample size needed in a within-participants design (N_W), relative to the sample size needed in a between-participants design (N_B), is (from Maxwell & Delaney, 2004, p. 562, Formula 47)

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

where a is the number of within-participants levels and ρ is the correlation between measurements in the population. This formula shows that switching from a between- to a within-participants design reduces the required sample size simply because each participant contributes data to each condition, even if the correlation between measurements is 0. In our example, a within-participants design would require only one third the number of participants as a between-participants design to achieve practically the same statistical power even when the three measurements are not correlated. Furthermore, a positive correlation would reduce the magnitude of the error term by removing systematic individual differences, and thereby increase the statistical power.

We can perform the simulation-based power analysis with the `ANOVA_power` or `ANOVA_exact` function:

```
power_within = ANOVA_power(design_within,
                           nsims = 100000)
exact_within = ANOVA_exact(design_within)
```

Recall that in our between-participants design, power was 81.14% when the enjoyment scores were uncorrelated. The power for a repeated measures ANOVA based on this design, when ratings for each of the three conditions are collected from 80 participants, is 98.38%. The effect size, as indicated by the results from the simulation, is much larger for the within-participants design (η_p^2 = .12) than for the three-group between-participants design (η_p^2 = .05). However, as explained by Olejnik and Algina (2003), it is difficult to compare η_p^2 across different research designs. Box 2 explains that the default calculation of η_p^2 by G*Power does not depend on the

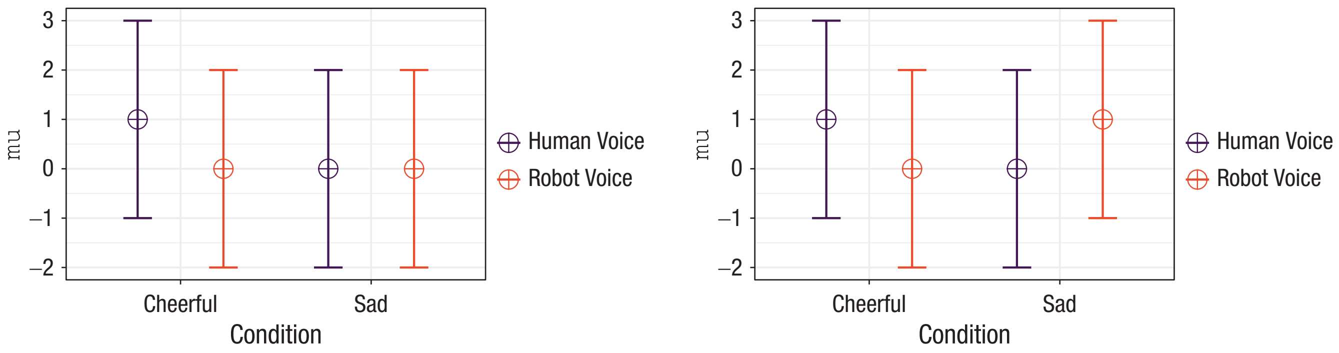


Fig. 5. Visualization of the expected means and standard deviations for an ordinal interaction (left) and a crossover (right) interaction in our hypothetical example. Error bars represent ±1 *SD*.

correlation among measures, and therefore differs from how other statistical software (including SPSS) calculates η_p^2. This peculiar choice for a default leads to errors for power analyses that include within-participants factors whenever researchers take a η_p^2 reported in the published literature and enter it in G*Power as the effect size (without changing the default power-calculation procedure by choosing the "as in SPSS" checkbox in the options menu). The *Superpower* package does not require researchers to enter η_p^2, but allows researchers to enter either a single value for the correlation between all dependent variables or a correlation matrix that specifies the expected population correlation for each pair of measurements.

Power for Interactions

So far, we have explored power analyses for one-factor designs. *Superpower* can easily provide statistical power for designs with up to three factors of up to 999 levels (e.g., `4b*2w*2w` would specify a mixed design with two within-participants factors each with two levels and one between-participants factor with four levels). Let us assume that we plan to perform a follow-up experiment in which, in addition to making the voice sound cheerful or sad, we introduce a second factor by making the voice sound more robotic compared to the default human-like voice, again collecting data from 80 participants in each condition. Different patterns of results could lead to observed interactions in this 2 × 2 design. Either no effect might be observed for the robotic voice or the effect observed for the robotic voice might be the opposite of that observed for the human-like voice (i.e., participants might enjoy a sad robotic voice more than a cheerful one, a "Marvin the Depressed Robot effect"). We specify the pattern of means as `(1, 0, 0, 0)` for the ordinal interaction, or as `(1, 0, 0, 1)` for the crossover (or dis-ordinal) interaction (see Fig. 5 for the expected pattern of means):

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

Simulations (using either the `ANOVA_power` or the `ANOVA_exact` function) show that we have 99.38% power for the crossover interaction when we collect data from 80 participants per condition and 60.62% power for the ordinal interaction. For comparison, the power for the simple effect comparing cheerful and sad human voices is 88.16%, similar to the power for the original one-way ANOVA we started with. Statistical power is much higher for the crossover interaction than for the ordinal interaction because the effect size is twice as large, as explained in Box 3. Statistical power is also higher for the crossover interaction than for the simple comparison, even though the effect size is identical (Cohen's $\hat{f}$ = 0.25), because the sample size has doubled. The interaction effect can be contrast-coded as 1, −1, −1, 1 to test the scores of 160 participants in the cheerful-human and sad-robot conditions against the scores of 160 participants in the cheerful-robot and sad-human conditions. The key insight here is that it is not the sample size per condition but rather the pooled sample size across conditions compared in a contrast that determines the power for the main effects and the interaction (cf. Westfall, 2015b).

Plotting Power Curves

The goal of an a priori power analysis is to determine the sample size required to reach a desired statistical power. By plotting the statistical power for each effect in an ANOVA design across a range of sample sizes, one can easily see which sample size would provide a

Box 3. Calculating Effect Sizes for the Interactions in the Hypothetical Example

Mathematically, the interaction effect is computed as the cell mean minus the sum of the grand mean, the marginal mean in each condition of one factor minus the grand mean, and the marginal mean in each condition for the other factor minus the grand mean (see Maxwell et al., 2017). For example, for the cheerful human-like voice condition in the crossover interaction, the calculation is 1 (the value in the cell) – (0.5 [the grand mean] + 0 [the marginal mean of cheerful voices minus the grand mean of 0.5] + 0 [the marginal mean of human-like voices minus the grand mean of 0.5]). Thus, the interaction effect is 0.5. Completing this calculation for all four cells for the crossover interaction gives the values 0.5, –0.5, –0.5, and 0.5. Cohen's *f* is then $\frac{\sqrt{\frac{0.5^2 + -0.5^2 + -0.5^2 + -0.5^2}{4}}}{2})$, or 0.25. For the ordinal interaction, the grand mean is (1 + 0 + 0 + 0)/4, or 0.25. Completing the calculation for all four cells for the ordinal interaction gives the values 0.25, −0.25, −0.25, and 0.25, and a Cohen's *f* of 0.125. Thus, the effect size of the crossover interaction is twice as large as the effect size of the ordinal interaction. Had we predicted a pattern of means of 2, 0, 0, 0, then the effect size for the ordinal interaction would have been 0.25. The take-home message is that a "medium" effect size (*f* = 0.25) translates into a much more extreme pattern of means in an ordinal interaction than in a dis-ordinal (crossover) interaction, or in a 2 × 2 × 2 interaction compared to a 2 × 2 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 than to perform a power analysis based on Cohen's *f* or η_p^2.

desired statistical power for all effects in the ANOVA. *Superpower* allows users to plot the statistical power across a range of sample sizes by specifying a desired statistical power and a maximum sample size. The plots will indicate if the desired power is reached for each effect, and if so, at which sample size. The code below specifies a 3 × 2 between-participants design (note that for two factors `a` and `b`, with three and two levels respectively, means are entered: `a1_b1`, `a1_b2`, `a2_b1`, `a2_b2`, `a3_b1`, `a3_b2`) and then calls the `plot_power` function to plot the power for designs with 10 to 100 participants per condition (see Fig. 6 for the power curve):

```
design_result <- ANOVA_design(
  design = "3b*2b", n = 50,
  mu = c(1, 2, 2, 3, 3, 4), sd = 3)
plot_power(design_result,
  min_n = 10, max_n = 100,
  desired_power = 90 , plot = TRUE)
```

There are two main effects, but no interaction effect. The main effect for factor `a` is the larger main effect, and 90% power is reached with 29 participants in each condition; for factor `b`, 90% power is reached with 64 participants in each condition. Because there is no interaction effect, only 5% Type I errors are expected for this effect, regardless of the sample size, and the desired power of 90% is never reached.

Plotting power curves across a range of sample sizes is implemented only for the `ANOVA_exact` function, and not for the `ANOVA_power` function because this is too resource intensive. Users of the latter function will need to steadily increase or decrease the sample size in their simulations to determine the sample size required to achieve the desired power for each effect.

Violation of Assumptions

So far, we have shown how simulations can be useful for power analyses for ANOVA designs when all assumptions of the statistical tests are met. An ANOVA is quite robust against violations of the normality assumption, which means the Type I error rate remains close to the alpha level specified in the test. Violations of the homogeneity-of-variances assumption can be more impactful, especially when sample sizes are unequal between conditions. When the equal-variances assumption is violated for a one-way ANOVA, Welch's *F* test is a good default (Delacre et al., 2019). When the sphericity assumption in within-participants designs is violated (when the variances of the differences between all pairs are not equal), a sphericity correction can be applied (e.g., the Greenhouse-Geisser or Huynh-Feldt correction) or a MANOVA can be performed. Alternative approaches for ANOVA designs with multiple between-participants factors include, for example, heteroscedasticity robust standard errors. *Superpower* allows researchers to perform power analyses in cases of unequal variances (or correlations) by performing Welch's *F* test, applying sphericity corrections, or a MANOVA.

Although some recommendations have been provided to assist researchers in choosing an approach to deal with violations of the homogeneity assumption (Algina & Keselman, 1997), it is often unclear if these violations of the homogeneity assumption are consequential for a given study. So far we have used simulations in *Superpower* to simulate patterns of means when

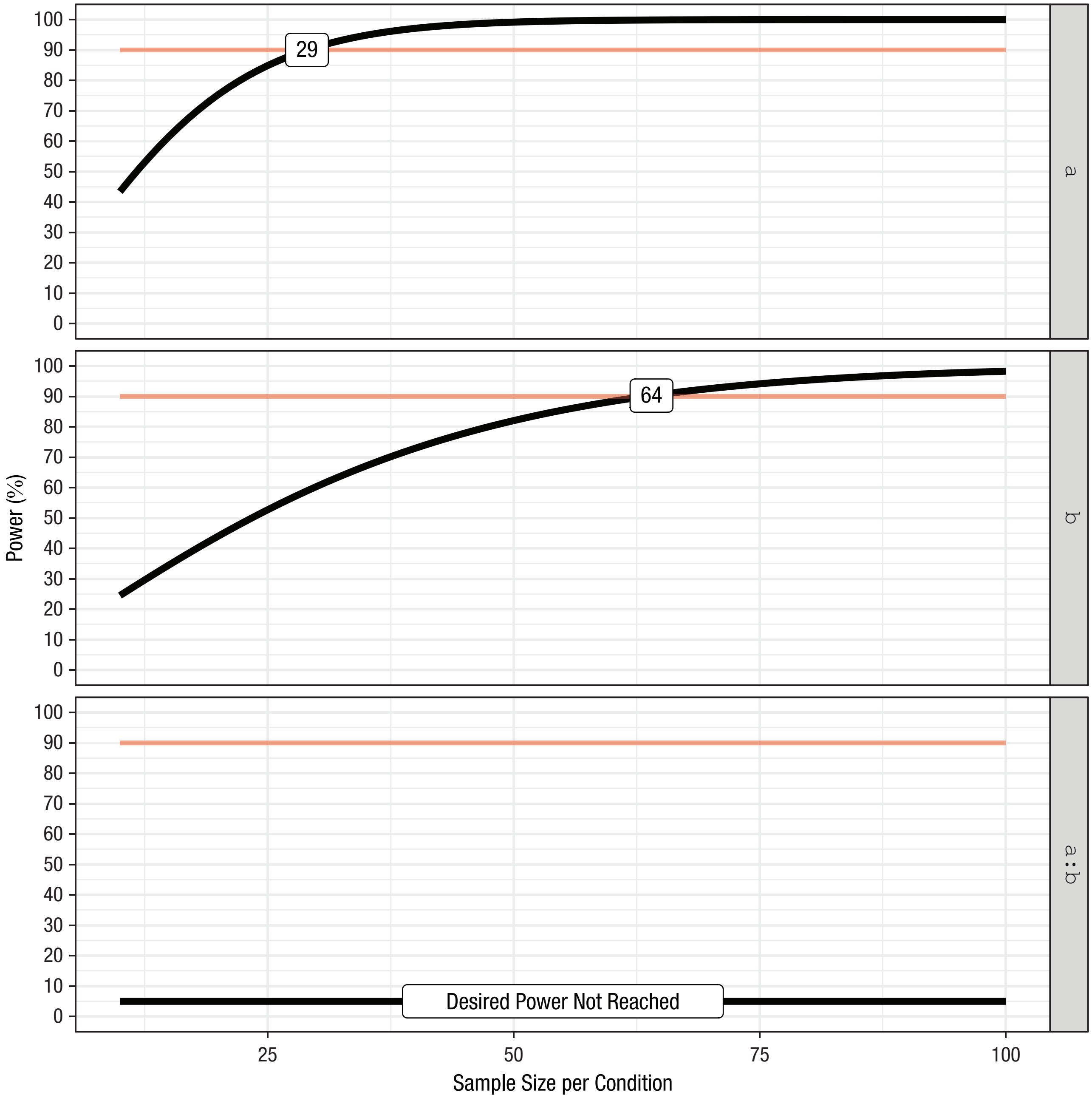


Fig. 6. Illustration of power curves across a range of sample sizes per group, from n = 10 to n = 100, for the two main effects and the interaction. The orange horizontal lines highlight the desired level of statistical power (90%).

there is a true effect, but we can also simulate a null effect. Such Monte Carlo simulation studies are used in published articles to examine the Type I error rate under a range of assumptions and for different tests. *Superpower* makes it easy to perform such simulation studies for the specific scenario a researcher is faced with, and can help a researcher decide whether violations of assumptions are something to worry about, and whether approaches chosen to deal with violations are sufficient.

As an example, let us revisit our earlier 2 × 2 between-participants design. Balanced designs (the same sample size in each condition) reduce the impact of violations of the homogeneity assumption, but let us assume that for some reason sample sizes varied between 20 and 80 per cell, and the population standard deviations varied extremely across conditions (from 1 to 5). We can use *Superpower* to estimate the impact of violating the homogeneity assumption by simulating a null effect (the means in all conditions are the same) and examining the

Type I error rate. We can specify a design with unequal sample sizes and unequal variances, as illustrated in the following code:

```
design_violation <- ANOVA_design(
  design = "2b*2b", n = c(20, 80, 40, 80),
  mu = c(0, 0, 0, 0), sd = c(3, 1, 5, 1),
  labelnames = c("condition",
                   "cheerful", "sad",
                   "voice",
                   "human", "robot"))
power_result = ANOVA_power(design_violation,
                              nsims = 100000)
```

This simulation indicates that the Type I error rates for the main effects and interactions in the ANOVA are approximately 15.85%. It is clear that the Type I error rate is too high. One solution would be to make sure that the experiment has equal sample sizes. If this is achieved, the Type I error rate is reduced to 4.98%, which is acceptable.

Conclusion

It is important to justify the sample size when designing a study. Researchers commonly find it challenging to perform power analyses for complex ANOVA designs that involve a mix of between- and within-participants factors. The R package, guide book, and Shiny apps (see https://arcaldwell49.github.io/SuperpowerBook, Caldwell et al., 2020) that accompany this article 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. The power for designs with specific patterns of means, standard deviations, and correlations between variables can be explored to choose a design and sample size that provides the highest statistical power for future studies. Simulation-based approaches can also help to provide a better understanding of the factors that influence the statistical power for factorial ANOVA designs or the impact of violations of assumptions on the Type I error rate.

Transparency

Action Editor: Daniel J. Simons
Editor: Daniel J. Simons
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.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

This work was funded by VIDI Grant 452-17-013 from the Netherlands Organisation for Scientific Research.

Open Practices

Open Data: not applicable
Open Materials: https://osf.io/pn8mc/
Preregistration: not applicable
All materials have been made publicly available via OSF and can be accessed at https://osf.io/pn8mc/. This article has received the badge for Open Materials. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

ORCID iD

Daniël Lakens https://orcid.org/0000-0002-0247-239X

Acknowledgments

Many improvements to *Superpower* are based on feedback from Lisa DeBruine and the `sim_design` function in her *faux* R package (DeBruine, 2020). The `ANOVA_exact` function was inspired by Chris Aberson's (2019) *pwr2ppl* package. We are grateful to Jonathon Love for proposing the name "Superpower" and developing a jamovi module.

The opinions or assertions contained herein are the private views of the author(s) and are not to be construed as official or reflecting the views of the Army or the Department of Defense. Any citations of commercial organizations and trade names in this report do not constitute an official Department of the Army endorsement of approval of the products or services of these organizations. Approved for public release; distribution is unlimited.

Notes

[1] For a detailed overview of the functionality of different software packages, see our supplemental file at https://osf.io/bwehv/.

[2] We refer to sample-level statistics (indicated with a hat) by default, and mention when we refer to population parameters instead.

References

Aberson, C. L. (2019). *Applied power analysis for the behavioral sciences* (2nd ed.). Routledge.

Algina, J., & Keselman, H. J. (1997). Detecting repeated measures effects with univariate and multivariate statistics. *Psychological Methods*, *2*(2), 208–218. https://doi.org/10.1037/1082-989X.2.2.208

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

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition*, *2*(1), Article 16. https://doi.org/10.5334/joc.72

Caldwell, A. R., Lakens, D., & Parlett-Pelleriti, C. M. (2020). *Power analysis with Superpower*. http://arcaldwell49.github.io/SuperpowerBook

Campbell, J. I. D., & Thompson, V. A. (2012). MorePower 6.0 for ANOVA with relational confidence intervals and

Bayesian analysis. *Behavior Research Methods*, *44*(4), 1255–1265. https://doi.org/10.3758/s13428-012-0186-0

Champely, S. (2020). *pwr: Basic functions for power analysis* (Version 1.3-0) [Computer software]. Comprehensive R Archive Network. https://CRAN.R-project.org/package=pwr

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.

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

DeBruine, L. (2020). *faux: Simulation for factorial designs* (Version 0.0.1.2) [Computer software]. Zenodo. http://doi.org/10.5281/zenodo.2669586

Delacre, M., Lakens, D., & Leys, C. (2017). Why psychologists should by default use Welch's *t*-test instead of Student's *t*-test. *International Review of Social Psychology*, *30*(1), 92–101. https://doi.org/10.5334/irsp.82

Delacre, M., Leys, C., Mora, Y. L., & Lakens, D. (2019). Taking parametric assumptions seriously: Arguments for the use of Welch's *F*-test instead of the classical *F*-test in one-way ANOVA. *International Review of Social Psychology*, *32*(1), Article 13. https://doi.org/10.5334/irsp.198

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

Haans, A. (2018). Contrast analysis: A tutorial. *Practical Assessment, Research & Evaluation*, *23*, Article 9. https://doi.org/10.7275/7dey-zd62

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. https://doi.org/10.1177/2515245918770963

Lang, A.-G. (2017). Is intermediately inspecting statistical data necessarily a bad research practice? *The Quantitative Methods for Psychology*, *13*(2), 127–140. https://doi.org/10.20982/tqmp.13.2.p127

Lenth, R. (2019). *emmeans: Estimated marginal means, aka least-squares means* (Version 1.4.8) [Computer software]. Comprehensive R Archive Network. https://CRAN.R-project.org/package=emmeans

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Erlbaum.

Maxwell, S. E., Delaney, H. D., & Kelley, K. (2017). *Designing experiments and analyzing data: A model comparison perspective* (3rd ed.). 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*, 537–563. https://doi.org/10.1146/annurev.psych.59.103006.093735

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods*, *8*(4), 434–447. https://doi.org/10.1037/1082-989X.8.4.434

Perugini, M., Gallucci, M., & Costantini, G. (2018). A practical primer to power analysis for simple experimental designs. *International Review of Social Psychology*, *31*(1), Article 20. https://doi.org/10.5334/irsp.181

Westfall, J. (2015a). *PANGEA: Power analysis for general anova designs*. Retrieved from at http://jakewestfall.org/publications/pangea.pdfwork?

Westfall, J. (2015b, May 26). 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/