

University of Groningen

## SampleSizePlanner

Kovacs, Marton; van Ravenzwaaij, Don; Hoekstra, Rink; Aczel, Balazs

*Published in:*  
Advances in Methods and Practices in Psychological Science

*DOI:*  
[10.1177/25152459211054059](https://doi.org/10.1177/25152459211054059)

**IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.**

*Document Version*  
Publisher's PDF, also known as Version of record

*Publication date:*  
2022

[Link to publication in University of Groningen/UMCG research database](#)

*Citation for published version (APA):*

Kovacs, M., van Ravenzwaaij, D., Hoekstra, R., & Aczel, B. (2022). SampleSizePlanner: A Tool to Estimate and Justify Sample Size for Two-Group Studies. *Advances in Methods and Practices in Psychological Science*, 5(1). <https://doi.org/10.1177/25152459211054059>

### Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

### Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

*Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.*

# SampleSizePlanner: A Tool to Estimate and Justify Sample Size for Two-Group Studies



Marton Kovacs<sup>1,2</sup>, Don van Ravenzwaaij<sup>3</sup>,  
 Rink Hoekstra<sup>4</sup>, and Balazs Aczel<sup>2</sup>

<sup>1</sup>Doctoral School of Psychology, ELTE Eotvos Lorand University, Budapest, Hungary; <sup>2</sup>Institute of Psychology, ELTE Eotvos Lorand University, Budapest, Hungary; <sup>3</sup>Department of Psychology, University of Groningen, Groningen, the Netherlands; and <sup>4</sup>Department of Educational Science, University of Groningen, Groningen, the Netherlands

Advances in Methods and  
 Practices in Psychological Science  
 January-March 2022, Vol. 5, No. 1,  
 pp. 1–10  
 © The Author(s) 2022  
 Article reuse guidelines:  
 sagepub.com/journals-permissions  
 DOI: 10.1177/25152459211054059  
 www.psychologicalscience.org/AMPPS



## Abstract

Planning sample size often requires researchers to identify a statistical technique and to make several choices during their calculations. Currently, there is a lack of clear guidelines for researchers to find and use the applicable procedure. In the present tutorial, we introduce a web app and R package that offer nine different procedures to determine and justify the sample size for independent two-group study designs. The application highlights the most important decision points for each procedure and suggests example justifications for them. The resulting sample-size report can serve as a template for preregistrations and manuscripts.

## Keywords

sample-size determination, power analysis, study design, open materials

Received 3/30/21; Revision accepted 9/25/21

Social and behavioral sciences are known to be plagued by undersampling (Ioannidis, 2005). In the traditional statistical framework, even when the effect exists, under-sampled studies yield either nonsignificant results or significant results because of overestimating the size of the effect. Because nonsignificant results are less likely to reach publications than significant ones, results of under-sampled studies either remain unpublished or impose a substantial bias on the body of published empirical findings. In addition, the low informational value of under-sampled studies may not justify the cost or potential risk they induce (Halpern et al., 2002). To mitigate these issues, authors are increasingly expected to plan and justify the sample size of their study (Maxwell, 2004). However, such sample-size justifications are meaningful only if they provide sufficient information to the readers to judge the adequacy of the author's decisions.

In the statistical literature, a few methods have been proposed to determine and justify sample size. In practice, however, authors are short of practical guides on how to navigate among the different sample-size methods. The aim of our tutorial is to point out for each

method the essential decision points that a researcher has to face during this process. We provide a short description of each method and the corresponding parameters, but we avoid listing their advantages and disadvantages. Because there are disagreements between the experts of the field regarding the correct use of some of the methods, we intentionally try to remain impartial and do not favor any of the presented methods. Researchers who want to know more about each method can find a number of useful references in the description of the methods. We also provide a collection of analysis code ready to use and a ShinyApp that helps researchers use and report the main sample-size-estimation techniques for different scenarios. The tutorial is focused exclusively on the scenario of the comparison of two independent groups (i.e., the independent *t*-test design) with a one-sided test. The most important terms related

## Corresponding Author:

Marton Kovacs, Doctoral School of Psychology, ELTE Eotvos Lorand University

Email: marton.balazs.kovacs@gmail.com



to sample size planning are defined in the glossary at the end of this article.<sup>1</sup>

## Sample-Size Determination and Justification

A lot of factors go into the determination of the sample size for an independent two-group study design. In this section, we first provide a bird's-eye view of the most important decisions. Next, we go into more detail on the specific inference tool that results from the combination of the larger choices.

It is crucial to not just state how one determined a planned sample size but to also give the reader insight into the reasons behind one's choices. In a recent overview, Lakens (2021) listed six types of general approaches to justify sample size in quantitative empirical studies: (a) measure entire population, (b) resource constraints, (c) a priori power analysis, (d) accuracy, (e) heuristics, and (f) no justification. For the first approach, no quantitative justification is necessary, and for the second approach, the researcher has no freedom to increase the sample size. Power analysis, or more generally, the estimation of true positive rate, is used when one plans to conduct hypothesis testing; accuracy justifications are used when one plans to conduct parameter estimation. Our tutorial mainly focuses on the resource constraints, a priori power analysis, and accuracy approaches and is aimed at providing a hands-on approach for the mechanical part of the sample-size determination (i.e., the calculation). For a deeper discussion of justification of these approaches or for other approaches (i.e., using heuristics or not providing justification), see Lakens (2021).

### Choosing a method in case of sample-size justification

In an ideal world, the choice for the number of participants would be solely determined by scientific considerations, and depending on the chosen technique, the collection of data would continue until either the desired sample size or a desired outcome has been reached. In practice, researchers are limited by time (collecting data is quite demanding), money (participants or people collecting the data may be paid, and the same may hold for renting space or equipment), or availability of participants (the population may be relatively small and/or the participation rate quite low).

When constrained by limited resources, it is important to be transparent about those limitations. It is also important to be open about scientific considerations. Depending on the nature of the study (Perhaps it is an initial exploration?), small sample sizes need not be a deal breaker. So although more data are always preferred from an informational point of view, by owning the limitations of a study, researchers improve future readers'

understanding of the process leading up to the eventual article and also answer in advance to readers who think the chosen sample size was insufficient.

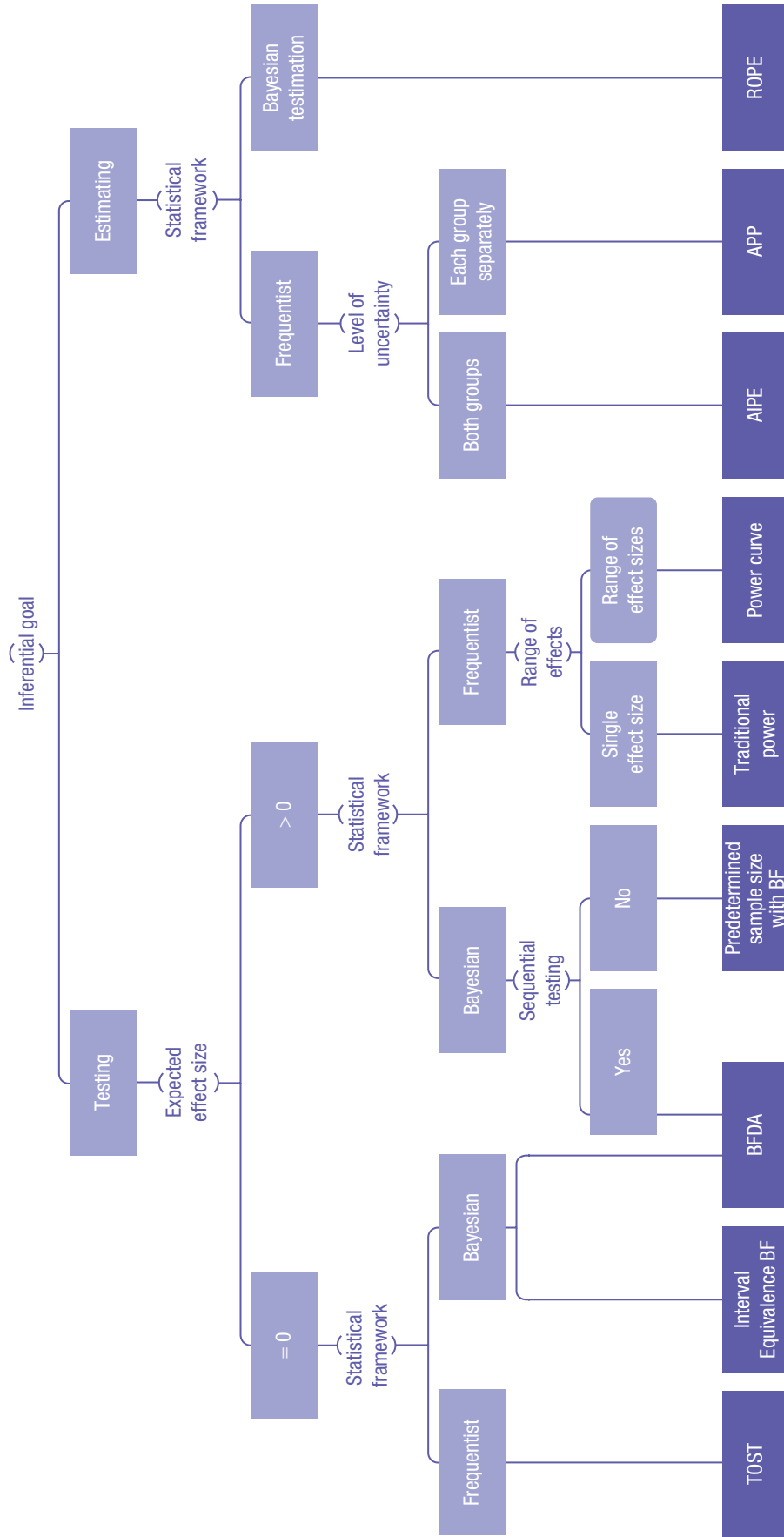
Whether or not authors have limited resources, two important choices need to be made: (a) whether they are interested in *statistical testing* or in *parameter estimation* and (b) whether they want to conduct their statistical inference within the *frequentist* framework or within the *Bayesian* framework. Starting with the first decision, statistical testing is the primary framework when one is interested in establishing whether an underlying population effect is equal to, different from, larger than, or smaller than a certain value. In essence, statistical testing lends itself to binary decision-making. Typically, testing is concerned with a fixed-point null hypothesis (e.g., there is no difference between two groups), although using intervals for testing is also possible. Alternatively, one might be interested in parameter estimation that is less interested in establishing the existence of a difference and instead is concerned with establishing the magnitude of the difference.

The second important decision concerns the statistical framework. Choosing to conduct statistical tests within a frequentist framework, one is usually interested in balancing the Type I (false positive) and Type II (false negative) error rates. Practitioners choosing to conduct statistical tests within a Bayesian framework are typically interested in being able to quantify the relative probability of hypotheses or models being true given the data and in including prior information.

Within the realm of statistical testing, there are some other factors that affect the preferred inference tool: Do you prefer to test for equivalence (no difference in mean) or for superiority (mean of one group larger than mean of other group), are you interested in calculating a required sample size for a specific hypothetical effect size or for a range of possible values, and do you wish to employ sequential testing (applicable to Bayesian testing)? In case of testing, some of the methods are designed to find support for the null hypothesis (e.g., two one-sided tests [TOST], region of practical equivalence [ROPE]), whereas others are designed to find support for the alternative hypothesis (e.g., traditional null hypothesis testing), and some methods are designed to find support for either (e.g., Bayes's factor design analysis [BFDA]). For frequentist estimation, the preferred inference tool might differ depending on whether one evaluated uncertainty for each group separately or jointly. We describe these specific factors when we go into detail about each of the preferred methods. A flow chart representing all of these choices is given in Figure 1.

### How to use this guide

In the next section, we illustrate the specific inference tools and resulting sample-size calculations in more



**Fig. 1.** The decisions that one faces when choosing among sample-size-estimation methods. The nine sample-size-estimation methods discussed in this article are listed in the bottom row. Some decisions are determined by the investigated question and the design of the study, whereas others are based on the preferred statistical framework.

detail using a ShinyApp and an R package we have developed. Throughout this section, we recurrently use two terms that have different meanings for different techniques. These are the *true positive rate* (TPR) and the *equivalence band* (EqBand). The TPR reflects the long-run probability of concluding there is an effect, given that it does exist. For traditional null hypothesis testing, this is typically referred to as *power*, but related concepts exist for different inference tools. The EqBand refers to an effect size region, typically around zero, that is deemed clinically insignificant or irrelevant. Different names are given to this region depending on the technique that employs them, such as statistical effect size of interest (SESOI) or ROPE. For both TPR and EqBand, we explain the specific meaning in context of the relevant inference tool below.

For each method, only the main parameters can be adjusted with a certain range of values in the ShinyApp by using a slider. These parameters are presented in the text in bold. Other parameters are set to preset values in the application but can be adjusted in the accompanying R package to any sensible value. These parameters are highlighted in italics in the tutorial. Both the app and the package allow the users to save or copy a text template with the results of the sample-size determination. We offer a list of possible justifications at the decision points for each method (indicated between brackets), but users are able to provide their own justification as free text. Note that the listed justifications are meant to provide guidance for the user, and they are not sufficient without further details provided by the researcher in the context of the given study. For example, previously reported values should always be accompanied by a theoretical justification of why these values make sense. The provided justification text could serve as a stub for the description of the chosen sample size in an article, a preregistration or registered report, or a grant proposal.

Throughout, we use the example story of Mary, the educational psychologist. Mary has come up with a new set of games that challenge spatial insight. She would like to test whether distributed and targeted engagement with these games for a period of 6 months for children in the age range of 8 to 12 will lead to lasting improvements on their IQ score as measured through Raven's progressive matrices test (population mean = 100,  $SD = 15$ ). Mary collects data for a control sample that gets regular education and for an experimental sample and plans to compare those samples. Mary has good reason to be skeptical about the effectiveness of training on increasing performance because there are several studies questioning the existence of such effects (Owen et al., 2010; Simons et al., 2016). For illustrative purposes, in some of the upcoming examples, Mary expects a null effect, and in others, Mary expects a positive effect

to highlight the different research scenarios for each sample-size-planning method. We also present a justification text for each sample-size-planning method based on Mary's choices described in the example research scenario for the given method.

The ShinyApp is available on <https://martonbalazs-kovacs.shinyapps.io/SampleSizePlanner>, and the R package can be installed by running the following command in R `devtools::install_github("marton-balazs-kovacs/SampleSizePlanner")`. There is more information about the R package and the ShinyApp at <https://github.com/marton-balazs-kovacs/SampleSizePlanner> or <https://marton-balazs-kovacs.github.io/SampleSizePlanner/>.

## Testing

### **Effect size = 0**

#### ***Two one-sided tests (TOST).***

*Study context.* Mary would like to know what sample size she needs for a power of .80 to study whether the mean IQ score of the experimental group's population is practically equivalent to the mean IQ score of the control group. She tests this assumption in a frequentist framework and considers a population effect size between  $-0.2$  and  $0.2$  to be "practically equivalent" to no difference. This would correspond to IQ scores between  $97 (100 + 15 \times -.2)$  and  $103 (100 + 15 \times .2)$ .

*Description.* TOST is a frequentist equivalence testing approach that adopts two one-sided hypotheses to designate an interval hypothesis (Schuirmann, 1987). The lower and upper boundaries of the interval are determined by the EqBand (i.e., SESOI) around the expected population effect size (e.g., 0). Lakens et al. (2018) listed several methods that can be used to determine the SESOI. In case of TOST, the two null hypotheses state that the effect size is equal to the lower and upper EqBand values, whereas the alternative hypotheses state that the effect size is significantly smaller than the upper EqBand value and significantly larger than the lower EqBandvalue. In case both one-sided tests reject the null hypothesis at a given significance level, the group means are considered to be practically equivalent. See Lakens et al. (2018) for further reading.

#### *Parameters.*

**Delta:** The expected population effect size. In most cases, this value will be zero.

**TPR:** The desired long-run probability of obtaining a significant result with TOST, given delta.

**EqBand:** The chosen width of the region for practical equivalence (i.e., the SESOI).

*Alpha:* The level of significance. The  $\alpha$  level in the application is preset to .05.



*How to use the package.* To use this method in R, run the following code: `SampleSizePlanner::ssp_tost(tpr = 0.8, eq_band = 0.2, delta = 0)`.

*How to report your sample-size estimation.* To calculate an appropriate sample size for testing whether the two groups are practically equivalent, we used the TOST (Schuirmann, 1987) method. We used an  $\alpha$  of .05. We set the aimed TPR to be 0.8 because [1] it is the common standard in the field; 2) it is the journal publishing requirement]. We consider all effect sizes below 0.2 equivalent to zero because [1] previous studies reported the choice of a similar equivalence band; 2) of the following substantive reasons: . . .]. The expected delta was 0 because [1] we expected no difference between the groups]. Given these parameters, a sample size of 429 per group was estimated to reach a TPR of 0.8 with our design.

### **Equivalence interval Bayes's factor.**

*Study context.* Mary would like to know what sample size she needs to have a long-run probability of .80 of obtaining a Bayes's factor (BF) larger than 10. Mary would like to test whether the mean IQ score of the experimental group's population is practically equivalent to the mean IQ score of the control group. Mary hypothesizes that there is no difference (i.e.,  $H_0$  is true). Mary tests this assumption in a Bayesian framework. Mary considers a population effect size between  $-0.2$  and under  $0.2$  to be practically equivalent. This would correspond to IQ scores between 97 ( $100 + 15 \times -.2$ ) and 103 ( $100 + 15 \times .2$ ).

*Description.* Equivalence interval BFs contrast an equivalence hypothesis to a nonequivalence hypothesis and quantify the evidence with BFs. Typically,  $H_0$  constitutes the equivalence interval (comparable with SESOI in the TOST framework), and  $H_a$  constitutes the complementary nonequivalence regions. Formally, the BF is calculated by dividing the fraction *posterior area inside the interval/posterior area outside the interval* (i.e., the posterior odds) by the fraction *prior area inside the interval/prior area outside the interval* (i.e., the prior odds). The resulting value quantifies how much more likely it is that the data occurred under a population effect size deemed equivalent relative to the data having occurred under a population effect size deemed nonequivalent. The current implementation uses a default Cauchy prior on effect size with the possible scale parameters of medium ( $r = 1/\sqrt{2}$ ), wide ( $r = 1$ ), or ultra-wide ( $r = \sqrt{2}$ ). For further reading, see Morey and Rouder (2011), Ravenzwaaij et al. (2019), and Linde et al. (2020).

#### *Parameters.*

**Delta:** The expected population effect size.

**TPR:** The desired long-run probability of obtaining a BF at least as high as the threshold, given delta.

**EqBand:** The chosen width of the equivalence region.

**PriorScale:** The scale of the Cauchy prior distribution. The PriorScale in the application can be set to  $1/\sqrt{2}$ , 1, and  $\sqrt{2}$ .

**Threshold:** Critical threshold for the BF. The threshold level in the application can be set to 10, 6, or 3.

*How to use the package.* To use this method in R, run the following code: `SampleSizePlanner::ssp_eq_bf(tpr = 0.8, delta = 0, eq_band = 0.2, thresh = 10, prior_scale = 1/sqrt(2))`.

*How to report your sample-size estimation.* To estimate the sample size, we used the interval equivalent BF (Morey & Rouder, 2011; Ravenzwaaij et al., 2019) method. We used a Cauchy prior distribution centered on 0 with a scale parameter of  $1/\sqrt{2}$ . We set the aimed TPR at 0.8 because [1] it is the common standard in the field; 2) it is the journal publishing requirement]. We consider all effect sizes below 0.2 equivalent to zero because [1] previous studies reported the choice of a similar equivalence region; 2) of the following substantive reasons: . . .]. The expected delta was 0 because [1] we expected no difference between the groups]. Our BF threshold for concluding equivalence was 10. Given these parameters, a minimal sample size of 144 per group was estimated to reach 0.8 TPR for our design.

### **Effect size >0 (frequentist)**

#### **Classical power analysis.**

*Study context.* Mary would like to know what sample size she needs for a power of .80 to study whether the mean IQ score of the experimental group's population is significantly higher than the mean IQ score of the control group. She tests this assumption in a frequentist framework for a hypothetical population effect size of 0.5. This corresponds to a mean IQ score of 107.5 in the experimental group ( $100 + 15 \times .5$ ), assuming a mean IQ score of 100 in the control group.

The classical power analysis approach allows one to calculate the required sample size to obtain a significant result for the null hypothesis test a certain proportion of times in the long run given an assumed population effect size.

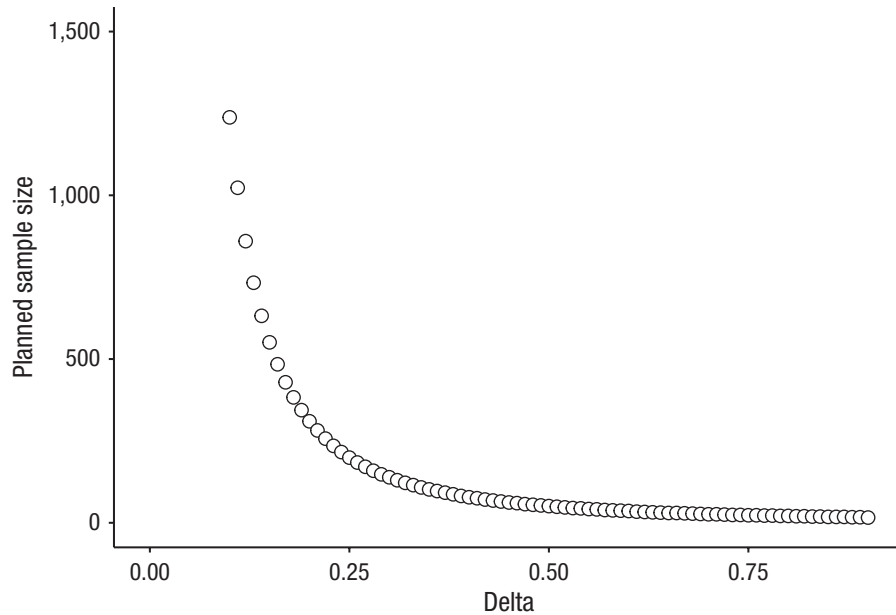
#### *Parameters.*

**Delta:** The expected population effect size.

**TPR:** The desired long-run probability of obtaining a significant result with a one-sided  $t$  test, given delta.

**Maximum N:** The maximum number of participants per group (both groups are assumed to have equal sample size).

**Alpha:** The level of significance. Alpha is preset to .05 in the application.



**Fig. 2.** The resulting power curve created by the application. The  $x$ -axis shows the range of deltas from the example, and the  $y$ -axis shows the corresponding sample sizes determined by the power curve method.

*How to use the package.* To use this method in R, run the following code: `SampleSizePlanner::ssp_power_traditional(tpr = 0.8, delta = 0.5, max_n = 5000, alpha = 0.05)`.

*How to report your sample-size estimation.* We used a power analysis to estimate the sample size. We used an  $\alpha$  of .05. We set the aimed TPR at 0.8 because [1] it is the common standard in the field; 2) it is the journal publishing requirement]. The expected delta was 0.5 because [1] previous results published in . . . ; 2) of the following substantive reasons: . . . ]. Given these parameters, a minimal sample size of 51 per group was estimated to reach 0.8 TPR for our design.

#### **Power curve.**

*Study context.* Mary would like to know what sample size she needs for a power of .80 to study whether the mean IQ score of the experimental group's population is significantly higher than the mean IQ score of the control group. She tests this assumption in a frequentist framework. However, she is reluctant to commit to a single hypothetical population effect size a priori, preferring to calculate required sample size for a range of hypothetical deltas between 0.1 and 0.9.

The power curve method is similar to a classical power analysis, but instead of calculating the appropriate sample size for one hypothesized population effect size, the method calculates the required sample size for a range of plausible population effect sizes.

*Parameters.*

**Delta:** A range of hypothetical population effect sizes.

**TPR:** The desired long-run probabilities of obtaining a significant result with a one-sided  $t$  test, given each value of delta.

**Maximum N:** The maximum number of participants per group (both groups are assumed to have equal sample size).

**Alpha:** The level of significance. Alpha is preset to .05 in the application.

*How to use the package.* To determine the sample sizes for each delta, see `curve_data <- SampleSizePlanner::ssp_power_curve(tpr = 0.8, delta = seq(0.1, 0.9, 0.01), max_n = 5000)`.

To plot the power curve, see `SampleSizePlanner::plot_power_curve(delta = curve_data$delta, n1 = curve_data$n1, animated = FALSE)`.

*How to report your sample-size estimation.* We used a power analysis to estimate the sample size. We used an  $\alpha$  of .05. We set the aimed TPR at 0.8 because [1] it is the common standard in the field; 2) it is the journal publishing requirement]. Because [1] we have no clear expectation of the magnitude of delta 2) we expected the delta to be around . . . ], we include power calculations for delta ranging from 0.1 to 0.9. Given these parameters, minimal sample sizes per group for different hypothetical effect sizes to reach 0.8 TPR can be found in Figure 2.

#### **Effect size >0 (Bayesian)**

##### **Predetermined sample size with Bayes's factor.**

*Study context.* Mary would like to test whether the mean IQ score of the experimental group's population

is higher than the mean IQ score of the control group. She would like to know what sample size she needs to have for a long-run probability of .80 of obtaining a BF larger than 10. Mary plans to collect all her data in one batch without testing sequentially. Mary expects the population effect size to be 0.5. This corresponds to a mean IQ score of 107.5 ( $100 + 15 \times .5$ ) in the experimental group, assuming a mean IQ score of 100 in the control group.

The present method calculates the corresponding default BF for a  $t$ -test statistic with Cauchy prior distribution centered on 0 with scale parameter of either  $1/\sqrt{2}$ , 1, or  $\sqrt{2}$  for several sample sizes (the so-called Jeffrey-Zellner-Siow BF; see e.g., Rouder et al., 2009). The function returns the optimal sample size needed to reach the TPR for a given BF threshold to detect an expected population effect size. If a range of possible population effect sizes are plausible under the given hypothesis, the function can calculate the optimal sample sizes for the given range of effect sizes and present the results in a figure (analogous to the Power Curve method). This method is designed to determine the sample sizes for the existence of an effect (i.e.,  $\delta > 0$ ).

#### Parameters.

**Delta:** The expected population effect size or a range of expected effect sizes.

**TPR:** The long-run probability of obtaining a BF at least as high as the critical threshold favoring superiority, given delta.

**Maximum N:** The maximum number of participants per group (both groups are assumed to have equal sample size).

**PriorScale:** The scale of the Cauchy prior distribution. The PriorScale in the application can be set to:  $1/\sqrt{2}$ , 1, and  $\sqrt{2}$ .

**Threshold:** Critical threshold for the BF. Three threshold levels are available in the app: 3, 6, and 10.

*How to use the package.* To use this method in R, run the following code: `SampleSizePlanner::ssp_bf_predetermined(tpr = 0.8, delta = 0.5, thresh = 10, max_n = 5000, prior_scale = 1/sqrt(2))`.

*How to report your sample-size estimation.* The following explains how to report your sample-size estimation: We used the Jeffrey-Zellner-Siow BF method to estimate the sample size. We used a Cauchy prior distribution centered on 0 with a scale parameter of  $1/\sqrt{2}$ . We set the aimed TPR at 0.8 because [1] it is the common standard in the field; 2) it is the journal publishing requirement]. The expected delta was 0.5 because [1] previous results published in . . . ; of the following substantive reasons: . . . ]. Our evidence threshold was 10. Given these parameters,

a minimal sample size of 105 per group was estimated to reach a 0.8 TPR for our design.

#### Bayes's factor design analysis (BFDA).

Mary would like to know what sample size she needs to have a long-run probability of .80 of obtaining a BF larger than 10. Mary would like to test whether the mean IQ score of the experimental group's population is higher than the mean IQ score of the control group in a Bayesian framework. Mary plans to collect all her data incrementally and thus is interested in using the advantage of not testing more than strictly necessary offered by sequential testing in her Bayesian analysis. Mary expects the population effect size to be 0.5. This corresponds to a mean IQ score of 107.5 in the experimental group ( $100 + 15 \times .5$ ), assuming a mean IQ score of 100 in the control group.

The description of the BFDA method is functionally identical to the one provided in the Predetermined Sample Size With BF section but gains in TPR because of the addition of sequential testing. In the app,  $H_0$  and  $H_a$  indicate the proportion of times sequential testing leads to BFs providing evidence with the given threshold for the null hypothesis and for the alternative hypothesis, respectively. Users of the Shiny app and R package should set delta to 0 if they wish to determine the sufficient sample size for rejecting an effect and use delta  $> 0$  if they wish to find support for the existence of an effect. For further reading, see Schönbrodt and Wagenmakers (2018) and Schönbrodt et al. (2017).

The parameters include the following:

**Delta:** The expected population effect size.

**TPR:** The long-run probability of obtaining a BF at least as high as the critical threshold favoring superiority, given delta.

**PriorScale:** The scale of the Cauchy prior distribution. The PriorScale in the application can be set to  $1/\sqrt{2}$ , 1, and  $\sqrt{2}$ .

**Threshold:** Critical threshold for the BF. Three threshold levels are available in the app: 3, 6, and 10.

To use this method in R, run the following code: `SampleSizePlanner::ssp_bfda(tpr = 0.8, delta = 0.5, thresh = 10, n_rep = 10000, prior_scale = 1/sqrt(2))`.

The following explains how to report your sample-size estimation: We used the BFDA method to estimate the sample size. We used a Cauchy prior distribution centered on 0 with a scale parameter of  $1/\sqrt{2}$ . We set the aimed TPR at 0.8 because [1] it is the common standard in the field; 2) it is the journal publishing requirement]. The expected delta was 0.5 because [1] previous results published in . . . ; 2) of the following substantive



reasons: . . . ]. Our evidence threshold was 10. Given these parameters, a minimal sample size of 81 per group was estimated to reach a 0.8 TPR for our design.

## Estimation

### Frequentist

#### Accuracy in parameter estimation (AIPE).

*Study context.* Mary would like to know what sample size she needs so that the 95% confidence interval for the population effect size has an expected width of 0.4. She estimates the population effect size to be 0.2.

*Description.* Accuracy in parameter estimation (AIPE) aims to determine the sufficient sample size to obtain a confidence interval with a desired width (precision) around the expected effect size (Kelley & Rausch, 2006). Note that the width of the calculated confidence interval will depend on the sample variance. As a result, it is possible that the variance is relatively large for a given sample, which leads to a resulting confidence interval that is larger than the width of the desired interval for a given sample. Thus, the AIPE method aims to establish the expected value of the calculated confidence interval, which can be thought of as the 50% long-run probability of obtaining a confidence interval no wider than the provided width.

*Parameters.*

**Delta:** The expected population effect size.

**Width:** The desired width of the confidence interval, given delta.

**Confidence level:** The desired level of confidence.

*How to use the package.* To use this method in R, run the following code: `SampleSizePlanner::ssp_aipe(delta = 0.5, width = 0.2, confidence_level = 0.8)`.

*How to report your sample-size estimation.* To estimate the sample size, we used the accuracy in parameter estimation [AIPE; Kelley and Rausch (2006)] method. We aimed for a 95% confidence level because [1] it is the common standard in the field; 2) it is the journal publishing requirement]. The desired width was 0.4 because [1] previous studies reported the choice of a similar region of practical equivalence; 2) of the following substantive reasons: . . . ]. We expected an underlying population effect size of 0.3 because [1] previous results published in . . . ; 2) of the following substantive reasons: . . . ]. Given these parameters, a minimal sample size of 195 per group was estimated for our design.

#### A priori precision (APP).

*Study context.* Mary would like to know the sample size for which she will have a 95% long-run probability

that the sample means in both the experimental and the control group lie within 0.2 *SD* (3 IQ points) of the true population mean.

*Description.* A priori precision (APP) aims to determine the sample size needed to have a certain long-run probability of both sample means being within a certain range of their respective population means, expressed in terms of standard deviations (Trafimow & MacDonald, 2017). As a result, APP is not reliant on the expected effect size.

*Parameters.*

**Closeness:** The desired closeness of the sample mean to the population mean defined in standard deviation.

**Confidence:** The desired probability of obtaining the sample mean with the desired closeness to the population mean.

*How to use the package.* To use this method in R, run the following code: `SampleSizePlanner::ssp_app(closeness = 0.2, confidence = 0.95)`.

*How to report your sample-size estimation.* To estimate the sample size, we used the a priori precision [APP; Trafimow and MacDonald (2017)] method. Before data collection, we wanted to be 95% confident that both sample means lie within 0.2 *SD* of the true population means. Given these parameters, the resulting minimum sample size was 126 per group for our design.

### Bayesian testimation

#### Region of practical equivalence (ROPE).

*Study context.* Mary would like to conduct parameter estimation to see whether the mean IQ score of her experimental group's population is practically equivalent to 100. She would like to know what sample size she needs to have a long-run probability of .80 of obtaining a 95% highest density interval (HDI) that is contained within her predefined ROPE. Mary hypothesizes that there is no difference (i.e.,  $H_0$  is true). She considers a population effect size between  $-0.2$  and under  $0.2$  to be practically equivalent. This would correspond to IQ scores between 97 ( $100 + 15 \times -0.2$ ) and 103 ( $100 + 15 \times 0.2$ ).

*Description.* The HDI-ROPE (often referred to as just ROPE) shares some features with the equivalence interval BF procedure. Both define an equivalence interval, construct a prior for the population effect size, and update to a posterior after the data come in. The equivalence interval BF procedure then focuses on the posterior and prior odds under complementary hypotheses. The ROPE procedure, on the other hand, identifies the 95% HDI (other percentages are permissible as well) and determines whether the HDI is fully

contained within the equivalence interval. For further reading, see Kruschke (2018) and Kruschke (2011).

#### Parameters.

**Delta:** The expected population effect size.

**TPR:** The desired long-run probability of having the HDI fully contained within the ROPE interval, given delta.

**EqBand:** The chosen ROPE interval.

**PriorScale:** The scale of the Cauchy prior distribution. The PriorScale in the application can be set to  $1/\sqrt{2}$ , 1, and  $\sqrt{2}$ .

*How to use the package.* To use this method in R, run the following code: `SampleSizePlanner::ssp_ropes(tps = 0.8, delta = 0.5, eq_band = 0.2, prior_scale = 1/sqrt(2))`.

*How to report your sample-size estimation.* To estimate the sample size, we used the region of practical equivalence (Kruschke, 2018) method. We used a Cauchy prior distribution centered on 0 with a scale parameter of  $1/\sqrt{2}$ . We set the aimed TPR at 0.8 because [1] it is the common standard in the field; 2) it is the journal publishing requirement]. We consider all effect sizes below 0.2 equivalent to zero because [1] previous studies reported the choice of a similar region of practical equivalence; 2) of the following substantive reasons: . . .]. The expected delta was 0 because [1] we expected no difference between the groups]. Given these parameters, a minimal sample size of 517 per group was estimated to reach a 0.8 TPR for our design.

## Summary

Justifying the decisions made during the sample-size planning process presents valuable information when one evaluates the inferences drawn from a study. The Shiny app and R package presented in this article aim to help researchers to choose and employ their sample-size estimation method. In addition, the tool provides assistance in reporting the process and justification behind sample-size choices. We encourage users and experts of the field to provide feedback and recommendations toward further developments.

## Transparency

*Action Editor:* Alexa Tullett

*Editor:* Daniel J. Simons

#### Author Contributions

M. Kovacs and D. van Ravenzwaaij are shared first authors. Conceptualization: M. Kovacs, D. van Ravenzwaaij, R. Hoekstra, and B. Aczel; methodology: D. van Ravenzwaaij; project administration: M. Kovacs; software: M. Kovacs and D. van Ravenzwaaij; supervision: B. Aczel; writing, original

draft preparation: M. Kovacs, D. van Ravenzwaaij, R. Hoekstra, and B. Aczel; writing, review and editing: M. Kovacs, D. van Ravenzwaaij, R. Hoekstra, and B. Aczel. All of the authors approved the final manuscript for submission.

#### 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.

#### Open Practices

Open Data: not applicable

Open Materials: <https://github.com/marton-balazs-kovacs/SampleSizePlanner/releases/tag/v0.1.0>

Preregistration: not applicable

All materials have been made publicly available via GitHub and can be accessed at <https://github.com/marton-balazs-kovacs/SampleSizePlanner/releases/tag/v0.1.0>. This article has received badges for Open Materials. More information about the Open Practices badges can be found at <http://www.psychologicalscience.org/publications/badges>.



## ORCID iDs

Marton Kovacs  <https://orcid.org/0000-0002-8142-8492>

Don van Ravenzwaaij  <https://orcid.org/0000-0002-5030-4091>

Rink Hoekstra  <https://orcid.org/0000-0002-1588-7527>

Balazs Aczel  <https://orcid.org/0000-0001-9364-4988>

## Note

1. A glossary of terms is provided below.

*Accuracy in parameter estimation (AIPE):* A sample-size-estimation method used for parameter estimation. The approach aims to find the required sample size such that the confidence interval has a certain expected width.

*A priori procedure (APP):* The approach aims to plan a sample size according to how close the researcher wishes both sample means to be to their respective population parameter and how confident the researcher wants to be in this.

*Bayesian inference:* A general framework for updating one's prior beliefs in light of new data.

*Bayes's factor design analysis (BFDA):* This technique provides an expected sample size such that compelling evidence in the form of a Bayes's factor can be collected for a given effect size with a certain long-run probability when allowing for sequential testing.

*Testing/estimation:* Two schools of inference focusing on establishing whether an effect exists as opposed to establishing the magnitude of an effect, respectively.

*Equivalence band (EqBand):* The region of effect sizes considered practically equivalent to zero. In our article, statistical effect size of interest (SESOI) and region of practical equivalence (ROPE) are subsumed under EqBand.

*Frequentist inference:* A general framework in which probabilities are defined as frequencies in hypothetical repeated events. In the context of statistical testing, frequentist inference is concerned with long-run error rates of rejecting the null hypothesis for the observed or more extreme parameters in a

given design when the model assumptions (e.g., independence of observations) are true.

*Statistical power:* The long-run probability of finding a significant effect given a certain population effect size.

*True positive rate (TPR):* The long-run probability of finding evidence for an effect, given that it exists. In our article, statistical power is subsumed under TPR.

*Classical power analysis:* This method is used to estimate the minimum sample size that a design needs to reach a certain level of statistical power given a desired significance level and expected effect size.

*Power curve:* This curve shows how changes in effect size modify the statistical power of a test.

*Region of practical equivalence (ROPE):* The region of effect sizes considered practically equivalent to zero under the highest density interval (HDI) ROPE method.

*Smallest effect size of interest (SESOI):* The region of effect sizes considered practically equivalent to zero under the TOST method.

*Sequential testing:* The practice of incrementally testing as data come in, typically until some predetermined level of evidence is obtained.

*Two one-sided tests (TOST):* A frequentist statistical testing approach aimed at establishing equivalence between two groups.

*Equivalence interval BF:* A Bayesian statistical testing approach aimed at establishing equivalence between two groups.

## References

- Halpern, S. D., Karlawish, J. H., & Berlin, J. A. (2002). The continuing unethical conduct of underpowered clinical trials. *JAMA*, *288*(3), 358–362.
- Ioannidis, J. P. (2005). Why most published research findings are false. *PLOS Medicine*, *2*(8), Article e124. <https://doi.org/10.1371/journal.pmed.0020124>
- Kelley, K., & Rausch, J. R. (2006). Sample size planning for the standardized mean difference: Accuracy in parameter estimation via narrow confidence intervals. *Psychological Methods*, *11*(4), 363–385. <https://doi.org/10.1037/1082-989X.11.4.363>
- Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, *6*(3), 299–312.
- Kruschke, J. K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science*, *1*(2), 270–280.
- Lakens, D. (2021). Sample size justification. *PxyArXiv*. <https://doi.org/10.31234/osf.io/9d3yf>
- 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.
- Linde, M., Tendeiro, J., Selker, R., Wagenmakers, E.-J., & van Ravenzwaaij, D. (2020). Decisions about equivalence: A comparison of TOST, HDI-ROPE, and the Bayes factor. *PxyArXiv*. <https://doi.org/10.31234/osf.io/bh8vu>
- Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, *9*(2), 147–163. <https://doi.org/10.1037/1082-989X.9.2.147>
- Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, *16*(4), 406–419. <https://doi.org/10.1037/a0024377>
- Owen, A. M., Hampshire, A., Grahm, J. A., Stenton, R., Dajani, S., Burns, A. S., Howard, R. J., & Ballard, C. G. (2010). Putting brain training to the test. *Nature*, *465*(7299), 775–778. <https://doi.org/10.1038/nature09042>
- Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237.
- Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, *25*(1), 128–142.
- Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*, *22*(2), 322–339. <https://doi.org/10.1037/met0000061>
- Schuurmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, *15*(6), 657–680.
- Simons, D. J., Boot, W. R., Charness, N., Gathercole, S. E., Chabris, C. F., Hambrick, D. Z., & Stine-Morrow, E. A. (2016). Do “brain-training” programs work? *Psychological Science in the Public Interest*, *17*(3), 103–186.
- Trafimow, D., & MacDonald, J. A. (2017). Performing inferential statistics prior to data collection. *Educational and Psychological Measurement*, *77*(2), 204–219.
- Van Ravenzwaaij, D., Monden, R., Tendeiro, J. N., & Ioannidis, J. P. (2019). Bayes factors for superiority, non-inferiority, and equivalence designs. *BMC Medical Research Methodology*, *19*(1), Article 71. <https://doi.org/10.1186/s12874-019-0699-7>