


**Method for Sample Size Determination for Cluster-Randomized Trials Using
the Bayes Factor**

Camila N. Barragán I. and Mirjam Moerbeek

Department of Methodology and Statistics, Utrecht University

Author Note

Camila N. Barragán I.  <https://orcid.org/00009-0001-5297-9746>

Correspondence concerning this article should be addressed to Camila N. Barragán I., Department of Methodology and Statistics, Utrecht University, Padualaan 14, 3584 CH Utrecht, The Netherlands. E-mail: c.n.barraganibanez@uu.nl

Abstract

Determining sample size is crucial in research study design. The hierarchical structure of the data in cluster-randomized trials (CRTs) complicates this process, thereby necessitating the determination of the sample size at each level. Most methods for these trials are based on null hypothesis significance testing (NHST), which has numerous pitfalls. These drawbacks can be avoided, however, by using the Bayes factor. Current methods for sample size determination when using the Bayes factor are limited to trials without multilevel structure. This study presents a method to determine the sample size for a one-period two-treatment parallel cluster-randomized trial using the Bayes factor. We introduce the implementation of this method in an R package. Simulation results show that the required sample size increases with decreasing effect sizes and with increasing intraclass correlation and Bayes factors.

Keywords: Bayes factor, Sample size determination, Cluster Randomized Trials, Sample size, Multilevel model

Method for Sample Size Determination for Cluster-Randomized Trials Using the Bayes Factor

Introduction

In the initial stages of the design of a research study, a key step is the determination of the sample size. Neglecting this key step may result in underpowered studies due to insufficient sample size, thereby potentially diminishing the ability to detect clinically relevant effects and leading to unethical use of participants. Furthermore, the publication of underpowered studies aggravates the crisis of replicability of research findings, as the replicability of a study is related to the statistical power of its design (Oakes, 1987). Determining sample size also prevents the use of more subjects than necessary, thereby reducing waste of resources and unethical participants recruitments.

Numerous elements come into play in determining the required sample size, with variations depending on the selected statistical model and the framework employed for hypothesis testing. The complexity of the interaction between the elements is especially intensified when dealing with multilevel models, given the hierarchical structure of the data. An example of multilevel data is found in cluster-randomized trials (CRTs), where complete groups, such as schools or families, are randomly assigned to treatment conditions. This design is widely used in social, behavioral, and biomedical sciences for the evaluation of treatments, programs, or interventions (Campbell & Walters, 2014; Donner & Klar, 2010; Eldridge & Kerry, 2012; Hayes & Moulton, 2009; Murray, 1998). Considering the hierarchical structure of the data, with subjects nested within clusters, the researcher must determine the required sample size for both levels, cluster sizes and number of clusters.

The conventional framework for hypothesis testing is based on null hypothesis significance testing (NHST), which is a combination of the significance testing approach of Fisher and the hypothesis testing approach of Neyman and Pearson (Balluerka et al., 2005). NHST involves the null hypothesis, i.e. the absence of an effect, and the alternative hypothesis, i.e. the presence of an effect. This approach to hypothesis testing assumes that

the null hypothesis is true in the population, and subsequently the researchers decide to either reject or retain the hypothesis using p-values. In the context of a CRT, previous studies have identified several factors that influence the determination of sample size within this framework, including intraclass correlation, effect size, Type I error rate, cluster size, and number of clusters (Moerbeek et al., 2000; Raudenbush, 1997a). Researchers determine the sample size using these elements in equations that illustrate the relation between sample size and power, or available software such as SPA-ML (Moerbeek & Teerenstra, 2016) and the Shiny CRT Calculator (Hemming & Kasza, n.d.).

An alternative approach to hypothesis testing is based on the Bayes factor. The Bayes factor is a quantification of the relative support of the data for one hypothesis over another (Heck et al., 2022; Hoijtink, 2012; Hoijtink, Mulder, et al., 2019; O'Hagan, 1995). In general, Bayesian sample size determination takes into account a user-specified minimum value of the Bayes factor and the effect sizes; however, various criteria exist in the literature (a comprehensive overview can be found in Gelfand and Wang, 2002). In the context of CRT, the methodology for determining the Bayesian sample size is scarce. Wilson (2022) proposed a method to calculate the total number of participants using Monte Carlo simulations, but this method was proposed for Bayesian inference instead of hypothesis testing, and assumes that the number of clusters is fixed beforehand, whereas in some CRTs, the cluster size is fixed beforehand, so that the number of clusters needs to be determined.

This study aims to present a method to determine the sample size in CRTs using the Bayesian approach for hypothesis testing. The method for sample size determination relies on simulation studies, for which we created functions in R that can either determine the required number of clusters given a fixed cluster size or, vice versa, the cluster size that is required for a fixed number of clusters. The next section introduces the data generation model. Subsequently, we discuss the shortcomings of NHST and the advantages of using the Bayes factor. We then explore the details of the determination of the sample size for

CRT, explaining each essential factor for sample size determination and the underlying algorithm. Subsequently, we present the results of a simulation and the sample size required for realistic scenarios. To conclude, we present the limitations of the methodology and offer suggested advice to researchers.

Cluster-Randomized Trials (CRTs)

The data from a CRT have a so-called multilevel structure, with variables measured on individuals at the first level and variables measured on clusters at the second level (Goldstein, 2011; Hox et al., 2017; Lazega & Snijders, 2016; Raudenbush & Bryk, 2010). An example of this design is the study by Ausems et al. (2002), in which the objective was to test the additional effect of out-of-school smoking prevention intervention. In this study, elementary schools were randomly assigned to four treatment conditions: an in-school smoking prevention program, a computer-based out-of-school smoking prevention program, a combined approach (in-school and out-of-school conditions), and a control condition. The students filled out a questionnaire twice, once before the intervention and once afterward. The researchers expected that students within the same school would mutually influence each other's smoking behavior; hence, the multilevel model was used to account for dependencies in the outcome variables.

Randomization at the cluster level rather than the individual level results in a decrease in statistical power, given the dependency on outcome measures within the same cluster. In other words, the CRT does not provide the same amount of information as an individually randomized trial. Despite this drawback, the CRT is widely used for ethical and logistical reasons. An advantage is that the design helps to avoid or reduce contamination of the control condition that may occur if multiple treatment conditions are available within each cluster, and information likely leaks from the intervention condition to the control. This leakage may occur when the intervention relies on providing new information, procedures, or guidelines to the participants (Moerbeek, 2005).

In this paper, the continuous outcome Y_{ij} for an individual i in cluster j , is a

function of the treatment condition:

$$Y_{ij} = \mu_C I_{Cj} + \mu_T I_{Tj} + u_j + e_{ij} \quad (1)$$

The μ represents the mean of the control condition (C) and the treatment condition (T).

The indicator variable I_{Cj} can take values 0 and 1 and indicates when cluster j is in the control condition, while I_{Tj} indicates when cluster j is in the treatment condition.

Additionally, two random terms are included, $u_j \sim N(0, \sigma_u^2)$ at the cluster level and $e_{ij} \sim N(0, \sigma_e^2)$ at the individual level, which are assumed to be independent of each other.

The sum of the between-cluster variance σ_u^2 and the within-cluster variance σ_e^2 results in the total variance, denoted as σ^2 . The two variances also define the intraclass correlation coefficient (ICC), $\rho = \sigma_u^2 / (\sigma_e^2 + \sigma_u^2)$, which is the proportion of total variance attributable to the cluster level.

The standardized treatment effect, denoted as δ and also known as the effect size, is defined as

$$\delta = \frac{\bar{\mu}_T - \bar{\mu}_C}{\sigma} \quad (2)$$

where σ is the standard deviation of the outcome variable. The variance of the treatment effect is expressed as

$$\frac{4\sigma^2[1 + (n_1 - 1)\rho]}{n_1 n_2} \quad (3)$$

Here, n_1 represents the number of individuals per cluster, and n_2 represents the total number of clusters.

All of these elements play a crucial role, along with statistical power, in the estimation of the sample size. The statistical power is denoted as $1 - \beta$, where β is the probability of committing a Type II error. In the context of CRT, the definition of

statistical power is given by the combination of Equations (2) and (3)

$$1 - \beta = \frac{\delta}{\sqrt{\frac{4\sigma^2[1+(n_1-1)\rho]}{n_1 n_2}}}. \quad (4)$$

This equation shows that the power decreases as ρ increases, especially when the common cluster size n_1 is high. Therefore, the researcher must balance the cluster sizes with the number of clusters to obtain the minimum sample required to detect a treatment effect.

One approach to determine the sample size is using the design effect

$$DE = 1 + (n_1 - 1)\rho \quad (5)$$

which considers the effect of clustering. The total number of subjects is calculated based on the sample size obtained for an individually randomized design and then is inflated by the design effect with the fixed cluster size n_1 (e.g., Campbell and Walters, 2014; Moerbeek and Teerenstra, 2016).

An alternative approach to determine the sample size uses the factors that influence power. Moerbeek and Teerenstra (2016) presented formulas describing the relation between statistical power, effect size, Type I error rate, ICC, and sample size. Equation (4) can be rewritten so that the number of clusters becomes a function of cluster size, Type I error rate, power, and effect size

$$n_2 = 4 \frac{1 + (n_1 - 1)\rho}{n_1} \left(\frac{z_{1-\alpha} + z_{1-\beta}}{\delta} \right)^2 \quad (6)$$

where z denotes the percentile from the standard normal distribution and α the significance level. The formula makes evident that increasing the common cluster size n_1 leads to a smaller number of clusters, while increasing the intraclass correlation ρ leads to a larger number of clusters. Alternatively, we can also formulate a function for the cluster size

$$n_1 = 4 \frac{1 - \rho}{\left(\frac{\delta}{z_{1-\alpha} + z_{1-\beta}} \right)^2 n_2 - 4\rho} \quad (7)$$

Here, it can be seen that, for a small number of clusters, the desired power level may not be achieved, even when the cluster size increases to infinity (Hemming et al., 2011).

However, important to note is that the methods for estimating the sample size discussed until now have been established for the NHST framework, which comes with notable limitations. These limitations will be explored in greater detail in the next section.

Hypothesis Testing

Null Hypothesis Significance Testing (NHST)

Despite the widespread use of null hypothesis significance testing (NHST), criticism of this approach has grown over the past few decades. Hooijink, Mulder, et al. (2019) provide an extensive account of numerous issues associated with NHST. One criticism in their paper is related to the use of the NHST approach in research. The excessive emphasis on p-values has contributed to publication bias, as studies yielding statistically significant results are more likely to be published. Furthermore, this emphasis on p-values has led some researchers aiming to advance their careers to engage in questionable research practices, such as p-hacking, hypothesizing after the results are known, and cherry-picking. A second criticism is that the use of a dichotomous decision rule based on $\alpha = 0.05$, or another appropriate value, narrows the focus of the investigation to reporting whether the null hypothesis is rejected.

A third criticism is the question whether one is even interested in testing the null hypothesis. The null hypothesis indicates that there is zero effect, or in other words, that two groups have exactly the same means on a continuous outcome variable. This hypothesis is operationalized as $H_0 : \mu_1 = \mu_2$, which means that the mean outcome in condition 1 is equal to the mean outcome in condition 2. However, the likelihood of this scenario in reality is very low, rendering the test practically unnecessary (Gu et al., 2014). A fourth criticism is that the NHST approach is focused on the null hypothesis, and when this hypothesis is rejected, the conclusion remains limited to asserting that the effect is not zero. When comparing more than two treatment means, post hoc tests are required to

understand which condition means significantly differ from each other.

Beyond Null Hypothesis Testing

Diverse types of hypotheses can be of interest to researchers; to illustrate some of those types, we consider the study presented in the "Cluster-Randomized Trials (CRTs)" section, where Ausems et al. (2002) collected data on four treatment conditions in a smoking prevention intervention. An informative hypothesis would show that there is an order between the treatment condition means. For instance,

$$H_1 : \mu_{combined} > \mu_{in} > \mu_{out} > \mu_{control}$$

where the mean of the combination of the in-school and out-of-school smoking prevention program is expected to be larger than the mean of the in-school program, which is larger than the mean of the computer-based out-of-school smoking prevention program, which in turn is larger than the mean of the control condition. Other informative hypotheses can contain different combinations between the following parameters:

$$H_2 : \mu_{combined} > \mu_{in}, \mu_{out} > \mu_{control}$$

$$H_3 : \mu_{combined} = \mu_{in}, \mu_{out} > \mu_{control}.$$

Under H_2 , the expectation is that the combined program has a mean greater than both in-school and out-of-school programs, with no prior expectation about the ordering of those in-school and out-of-school means. However, both of these means are anticipated to be higher than the control condition mean. The hypothesis H_3 states that the combined program mean is equal to the in-school mean, and both means for in-school and out-of-school program conditions surpass the control condition mean. These hypotheses use equality and inequality constraints to specify the relations between the treatment means and are therefore called constrained hypotheses. On the other hand, a hypothesis without any constraint is called unconstrained

$$H_u : \mu_{combined}, \mu_{in}, \mu_{out}, \mu_{control}$$

Such a hypothesis implies that the researcher does not have any a priori expectations concerning the group means. A final hypothesis that is used in this paper is the complement hypothesis, which in the case of H_1 the complement is

$$H_c : \mu_{combined} < \mu_{in} < \mu_{out} < \mu_{control}$$

This hypothesis covers the space of all possible parameter values that are not covered in H_1 .

Important to note is that, in this paper, we use only two treatment conditions: treatment and control. Moreover, recognizing the broad use of the null hypothesis in social sciences, as well as considering that the project aims to provide researchers with the necessary tools, we included the hypothesis with equality constraint as a possible hypothesis of the study. The pair of competing hypotheses, including the equality constraint, is referred to as Hypotheses Set 1:

$$H_0 : \mu_C = \mu_T$$

$$H_1 : \mu_C < \mu_T$$

while Hypotheses Set 2 contains a pair of informative hypotheses:

$$H_1 : \mu_C < \mu_T$$

$$H_2 : \mu_C > \mu_T$$

This set of hypotheses can be used when one has good reason to believe the treatment is performing better than the control (H_1) and one wants to test this versus the opposite (H_2).

Bayes Factor

The Bayes factor is a quantification of the relative support of the data for one hypothesis over another (Heck et al., 2022). It is also known as the ratio of two marginal

likelihoods, or the marginal probability of the data X under H_i or $H_{i'}$ (Heck et al., 2022; Kass & Raftery, 1995). This quantification is represented as

$$BF_{ii'} = \frac{P(X|H_i)}{P(X|H_{i'})}$$

However, when comparing a constrained hypothesis with an unconstrained one, the formulation can be simplified by using the so-called encompassing prior approach, where the constrained hypothesis is nested within the unconstrained one (Gu et al., 2018). Thus, the evaluation of the null hypothesis and an informative hypothesis is calculated by

$$BF_{0i} = \frac{BF_{0u}}{BF_{iu}} = \frac{\frac{f_0}{c_0}}{\frac{f_i}{c_i}} \quad (8)$$

where f is the relative fit and c is the relative complexity of the hypothesis under consideration compared to the unconstrained hypothesis. The fit can be interpreted as the proportion of the posterior distribution that is supported by the informative hypothesis (H_i), and the complexity is the proportion of the prior distribution supported by the informative hypothesis (H_i) (Klugkist et al., 2005). However, the computation of the fit and complexity of the null hypothesis (H_0) is carried out by using the density instead of the proportion; in this case, the density of the difference between the means in treatment and control condition is zero.

In the case of evaluating two informative hypotheses, the Bayes factor is calculated by

$$BF_{ic} = \frac{BF_{iu}}{BF_{cu}} = \frac{\frac{f_i}{c_i}}{\frac{1-f_i}{1-c_i}} \quad (9)$$

where the Bayes factor is the ratio of the Bayes factors for the informative hypothesis (H_i) compared to the unconstrained hypothesis (H_u) and the Bayes factor of the complement hypothesis (H_c) compared to H_u .

In the encompassing prior approach, the prior distribution is constructed using a truncation of the unconstrained hypothesis (Klugkist et al., 2005). According to this

approach, the prior distribution is specified must meet two characteristics. One, the prior must be neutral and not favor any hypothesis, which means that the prior distribution is the same for both hypotheses; so, for Hypotheses Set 2, the prior is set to 0.5. Two, the prior should be non-informative, which means that the prior is specified such that its variance is large enough to be vague and for the data to dominate in the computation of the Bayes factor (Klugkist & Hoijtink, 2007). In this paper, we use the approximated adjusted fractional Bayes factor (AAFBF). Along with employing the encompassing prior approach, AAFBF specifies the prior distribution, using a fraction of the information in the data, denoted as b . Additionally, the fractional prior is centered around a focal point, ensuring that no hypothesis under consideration is favored. Due to the large sample theory, both the marginal posterior and the fractional prior can be approximated, using a normal distribution with mean in the maximum likelihood estimate and a covariance matrix Σ for the posterior and $\frac{\Sigma}{b}$ for the prior. To conclude, the prior distribution is normally approximated, centered around zero, and the variance is determined using a fraction of the data. Further details can be found in Appendix A.

The approximate adjusted fractional Bayes factor is defined as

$$AAFBF_{iu} = \frac{\int_{\theta \in \Theta_i} \pi_u(\theta|X) d\theta}{\int_{\theta \in \Theta_i} \pi_u^*(\theta|X^b) d\theta} \quad (10)$$

where the numerator is the posterior distribution of the parameter of interest (θ) and the denominator is the adjusted fractional prior distribution. The parameter space in Θ_i represents the parameter space in agreement with the hypothesis H_i .

The AAFBF is sensitive to the parameter b when evaluating a hypothesis with equality constraints; it is therefore crucial to perform a sensitivity analysis with different choices of fractions of information b (Gu et al., 2018). In the evaluation of a hypothesis that only includes inequality constraints, AAFBF is stable regardless of the fraction of information b (Mulder, 2014). Further explanation of b can be found in Appendix A.

One of the advantages of the Bayes factor is that it is easy to interpret. Given that

the Bayes factor is a quantification of the support for one hypothesis over the other, the interpretation is how much relative support the data have for one hypothesis over the other. For instance, if $BF_{10} = 10$ then the relative support for hypothesis 1 (H_1) is ten times larger than for the null hypothesis (H_0). In the case where the Bayes factor is 1, there is no preference between the hypotheses under consideration. Important to note is that the Bayes factors can take on positive values to infinite. Initially, Jeffreys (1983) proposed a threshold of 3.2 to declare that there is “positive” evidence for one hypothesis. Later, Kass and Raftery (1995) proposed more thresholds to distinguish between positive, strong, and very strong evidence in favor of one hypothesis. Nevertheless, another advantage of the Bayes factors is that there are no strict thresholds because their interpretation is relative to the hypotheses of study; for this reason, avoiding the use of cut-off values for interpretation is strongly advised.

Methodology for Sample Size Determination

Fu (2022) proposed a method to determine the sample size using the Bayes factor for hypothesis testing. The sample size is determined by the probability (η) that the Bayes factor exceeds a threshold (BF_{thresh}), given that the hypothesis is true. This probability is,

$$P(BF_{ii'} > BF_{thresh} | H_i) \geq \eta \quad (11)$$

where i and i' represent competing hypotheses of Hypotheses Set 1 or Hypotheses Set 2. Equation (11) is evaluated for each of the hypotheses under consideration (i.e. $BF_{ii'}$ and $BF_{i'i}$). The probability and the threshold are specified by the researcher taking into account the purpose of the study; in cases where the study is high-stakes and the aim is to obtain compelling evidence with a large probability, the threshold and the probability are relatively large. Consider as an example a study involving the evaluation of the effectiveness of a public health intervention in reducing the transmission of diseases spread by mosquitoes; in this case, the researcher may choose to adopt a high threshold of 10 and a probability of 0.9. However, when the aim is to collect evidence for exploratory studies,

researchers may choose a lower probability and threshold. To illustrate, picture a study where the aim is to compare the effect of using cognitive-behavioral therapy alone with the effect of using an app with AI, developed to interact with patients with anxiety disorders, as a complement of cognitive-behavioral therapy. Considering that the study is exploratory, the researcher may use a low threshold of, say, 3 and with a probability of 0.8.

As mentioned earlier, for a CRT, two sample sizes need to be determined: cluster sizes and number of clusters. The strategy proposed in this paper fixes one of the samples, with the other to be determined. The algorithm can be seen in Figure 1.

The first step in this method is to generate the data sets corresponding to a two-group-parallel conditions cluster-randomized trial, with a given number of clusters and cluster size. The second step is to fit the multilevel model in Equation (1) to the data with the function `lmer` from the R package `lme4` (Bates et al., 2015). The third step is to use the estimates of means and variance of the means for both treatment conditions to calculate the Bayes factors for both hypotheses in Hypotheses Set 1 or Hypotheses Set 2. The fourth step is to calculate the proportion of generated data sets for which the Bayes factors exceed the threshold and to evaluate the Bayesian power criterion. The fifth step, which occurs when the power criterion in Equation (11) is not met, is to change the sample size. Rather than increasing the sample size by only one, the algorithm incorporates a binary search to efficiently find the required sample size and reduce the computation time (see Appendix B). In the case that one of the hypotheses contains only equality constraints, a sensitivity analysis is carried out, which means that the aforementioned first five steps are repeated for different choices of fraction of information b as specified by the researchers. When the power criterion is met and the sensitivity analysis has finished, the results are displayed in a table where the researcher can find the hypotheses under consideration, the number of clusters, the cluster sizes, and the probabilities of Bayes factors exceeding the threshold.

In the repository and in the R package, two functions to determine the sample size for a trial with two parallel treatment conditions can be found. The function

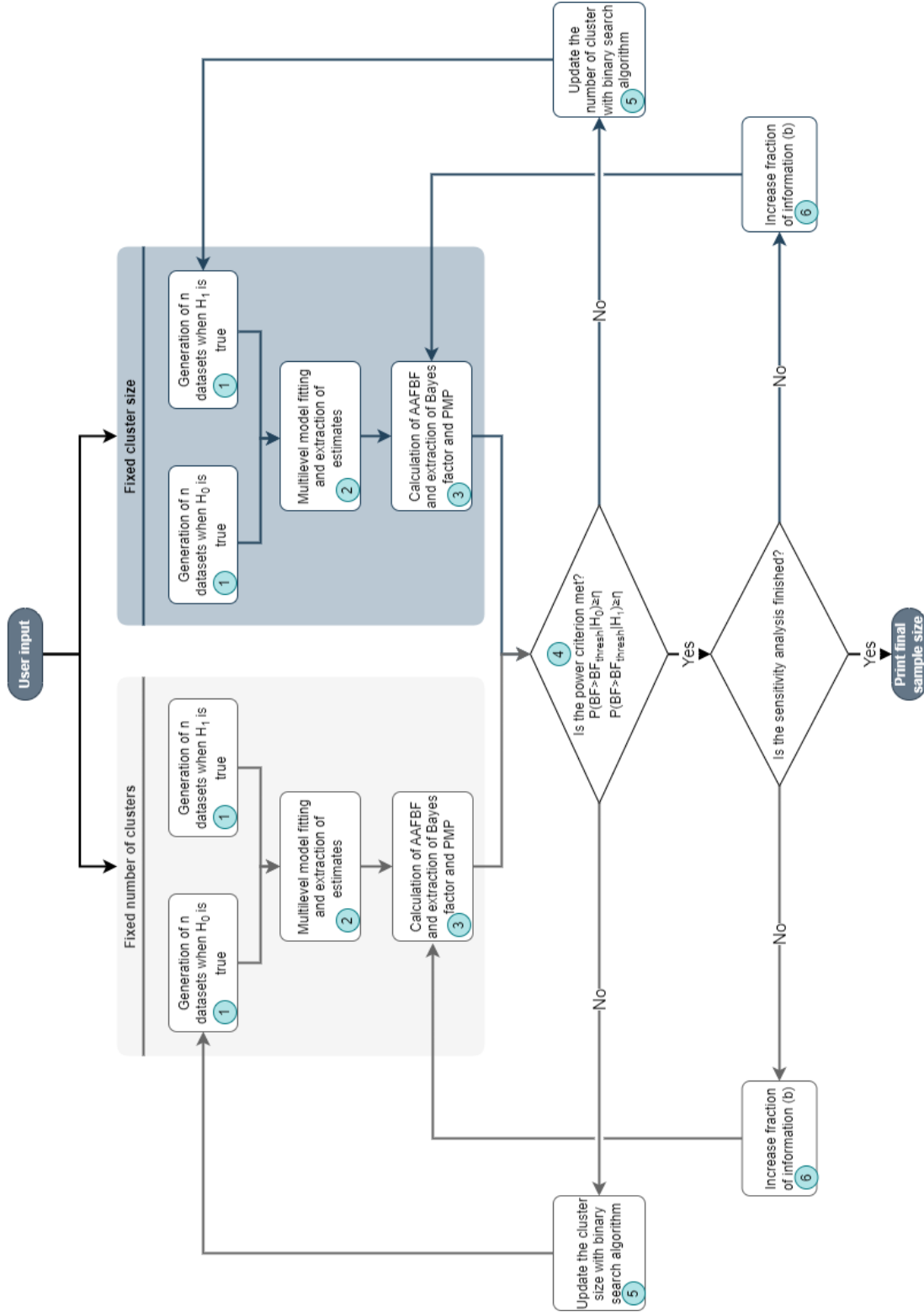


Figure 1

Algorithm of function for sample size determination in cluster-randomized trials when using the Bayes factor to test informative hypotheses, including the null hypothesis.

`SSD_crt_null` determines the sample size when one of the hypotheses has an equality constraint, which is the Hypotheses Set 1. Meanwhile, the function `SSD_crt_inform` determines the sample size for Hypotheses Set 2.

The arguments necessary to determine the sample size are the following:

- **eff_size** is a numeric value corresponding to the standardized mean difference between the treatment and control conditions. The effect size is the standardized difference between the treatment and control conditions.
- **n1** is a numeric value that specifies the sizes of the clusters. All clusters are assumed to have the same sizes. The default value is 15 individuals in each cluster.
- **n2** is a numeric value that specifies the total number of clusters; as the same number of clusters is assigned to both treatment conditions, the value must be even. The default is 30 clusters: 15 in the experimental condition and 15 in the control condition.
- **BF.thresh** is a numeric value that specifies the desired minimum of the Bayes factor. This value indicates how much support the data should show for one of the hypotheses under consideration. The default value is 3.
- **eta** is a numeric value that indicates the probability of finding a Bayes factor equal or larger to the threshold, given that the hypothesis is true. The default value is 0.8.
- **ndatasets** is a numeric value that indicates how many data sets are generated to evaluate the power criterion. The default is 5,000 data sets.
- **rho** is a numeric value that specifies the intraclass correlation.
- **fixed** is a string that specifies which sample size is fixed. When the number of clusters is fixed (**fixed**="n2"), the function determines the cluster sizes. If the cluster sizes are fixed (**fixed**="n1"), then the function determines the number of clusters. The default setting is "n2".

- **max** is a numeric value that indicates the maximum sample size that is used by the algorithm: if the algorithm reaches this sample size, it stops. By default, the maximum sample size is 1,000.
- **batch_size** is a numeric value that indicates the batch size in the multilevel model fitting, which is a strategy to improve memory usage efficiency and computational performance, given that the data sets might become very large and require a considerable amount of computational effort for model fitting. The default is 100 models at the same time.

The function `SSD_crt_null` has one additional argument:

- **b_fract** is a numeric value that specifies the maximum value that the fraction of information b is multiplied by in the sensitivity analysis. A sensitivity analysis is carried out for all integer values ranging from 1 to b_{fract} . This means, the fraction of information taken from the data increases from $1/N_{eff}$ until and including $b_{fract} \times 1/N_{eff}$. For further information, the reader can refer to Appendix A. By default, **b_fract** is equal to 3.

The outputs are different for `SSD_crt_null` and `SSD_crt_inform`. However, for both functions, the output includes the hypotheses under consideration, the sample size required, whether the number of clusters or the cluster size was fixed, the probabilities that the Bayes factor is higher than the threshold, and data sets containing the Bayes factors calculated during the simulation. For `SSD_crt_null`, the output also incorporates the results for different choices of b .

Simulation Study

Design

Four simulations were carried out to provide sample sizes for various realistic scenarios. Two of the simulations had the objective of determining the number of clusters

given a fixed cluster size, while the other two aimed to determine the cluster size for a fixed number of clusters. The common factors in the design were the following:

- Intraclass correlation: 0.025, 0.05, 0.1
- Effect size: 0.2, 0.5, 0.8
- Bayes factor threshold: 1, 3, 5
- Probability: 0.8
- Maximum sample size: 1,000
- Eta: 0.8

Determining the number of clusters

To determine the number of clusters, the cluster size was fixed to the following values:

- Cluster sizes: 5, 10, 40

Together with the factors that were common for both simulations, 81 combinations were formed. For each of these combinations, 5,000 data sets were generated.

Determining the cluster size

To determine the cluster sizes, the number of clusters was fixed to the following values:

- Number of clusters: 30, 60, 90

The total of formed combinations of the factors was 81, and for each of these combinations, 5,000 data sets were generated.

The minimum sample size was set inside the functions; for cluster size was 5, and for number of clusters was 6, whereas the maximum was set to 1,000 in the functions' argument. The reason for using a maximum sample size was that, for a small number of

clusters, sufficient power is not always achieved in the framework of NHST, even in the case in which the cluster size increases to infinity (Hemming et al., 2011). In addition, for testing Hypothesis Set 1, a sensitivity analysis was carried out for each combination with fractions of information b , $2b$, and $3b$.

Results

Taking into account the limited space, this subsection presents a selection of the required sample sizes in tables. Readers interested in exploring additional results not displayed here can access them in the following Shiny app (<https://utrecht-university.shinyapps.io/BayesSampleSizeDet-CRT/>).

Determining the number of clusters for Hypotheses Set 1

Table 1 presents the required number of clusters and, for hypotheses H_0 and H_1 , the probability η of exceeding the threshold BF_{thresh} . For instance, for $BF_{thresh} = 1$, $n_1 = 5$, $ICC = 0.025$ and effect size $\delta = 0.2$ the required number of clusters is 126. This number results in probability $\eta = 0.807$ for H_1 and $\eta = 0.982$ for H_0 . Thus, in this specific case, the desired threshold is exceeded more often for H_0 than it is for H_1 . However, that is not necessarily the case for all other combinations of design factors in Table 1.

The results in Table 1 show that the required number of clusters to meet the power criterion increases as the ICC increases. This increase is expected, given that the higher the correlation between the individuals within a cluster, the larger the dependency among the individuals and, hence, the lower the effective sample size (Hox et al., 2017). This expectation is also possible to infer from Equation (6). Table 1 further shows a trade-off between the two sample sizes: the required number of clusters decreases if the cluster size increases. This decrease is obvious, given that the larger the cluster size, the more information is available within each cluster and, hence, fewer clusters are needed. The results also show an inverse relationship between the effect size and the number of clusters. This inverse result is because larger effect sizes are easier to detect than smaller effect sizes. The required number of clusters increases with the Bayes factor threshold. Increasing the

Table 1*Required number of clusters per treatment condition with a fraction of information equal to 1 and probability (η) of 0.8 for**Hypotheses Set 1*

		ICC = 0.025				ICC = 0.050				ICC = 0.100			
		Eff. Size = 0.2		Eff. Size = 0.5		Eff. Size = 0.2		Eff. Size = 0.5		Eff. Size = 0.2		Eff. Size = 0.5	
BF_{thresh}	n_1	Hypothesis	n_2	$P(BF > BF_{thresh})$	n_2	$P(BF > BF_{thresh})$	n_2	$P(BF > BF_{thresh})$	n_2	$P(BF > BF_{thresh})$	n_2	$P(BF > BF_{thresh})$	n_2
1	5	H0	126	0.982	16	0.942	138	0.982	18	0.941	158	0.981	20
		H1		0.807		0.808		0.810		0.820		0.806	
1	10	H0	72	0.984	12	0.943	80	0.980	14	0.942	110	0.982	14
		H1		0.816		0.891		0.800		0.896		0.801	
1	40	H0	30	0.981	8	0.943	42	0.977	8	0.927	70	0.983	10
		H1		0.821		0.976		0.802		0.913		0.804	
3	5	H0	160	0.944	26	0.813	174	0.944	28	0.802	204	0.946	32
		H1		0.804		0.872		0.808		0.873		0.801	
3	10	H0	92	0.940	14	0.806	110	0.943	18	0.809	140	0.945	24
		H1		0.810		0.857		0.813		0.887		0.802	
3	40	H0	38	0.937	8	0.822	54	0.940	10	0.809	90	0.943	16
		H1		0.828		0.941		0.803		0.903		0.800	
5	5	H0	176	0.904	68	0.802	188	0.898	76	0.804	228	0.905	88
		H1		0.808		0.998		0.800		0.998		0.805	
5	10	H0	98	0.892	42	0.811	118	0.897	48	0.802	154	0.900	58
		H1		0.801		0.999		0.801		0.999		0.811	
5	40	H0	40	0.889	16	0.804	60	0.898	24	0.807	100	0.896	42
		H1		0.805		0.998		0.810		0.999		0.802	

 n_1 represents the cluster size and n_2 represents the number of clusters.

threshold means that a larger number of clusters is needed to achieve the support for the correct hypothesis.

As the reader can easily verify from our Shiny app (<https://utrecht-university.shinyapps.io/BayesSamplSizeDet-CRT/>), overall, the number of clusters required increases with the fraction of information b , especially when the effect size is small. The explanation of this finding can be found in Fu et al. (2021), who indicated that when b gets larger, the prior variances decrease, while the complexity c_0 increases in Equation (8), resulting in smaller values of the Bayes factor when H_0 is true.

Determining the cluster sizes for Hypotheses Set 1

Table 2 presents the required cluster sizes as a function of the number of clusters, effect size, ICC, and Bayes factor thresholds. This table is similar in format to Table 1, but now the number of clusters rather than the cluster size appears in the second column, while the other columns show the required cluster sizes and corresponding Bayesian power.

As shown in Table 2, the desired power $\eta = 0.8$ for both hypotheses was not always achieved. The maximum specified cluster size of 1,000 was reached for one or both hypotheses without meeting the power criterion, especially in cases of small effect sizes and medium-to-high ICC ($\rho = 0.05$ and $\rho = 0.1$, respectively). This outcome was expected, given that an increase in the number of individuals only has a limited increase in power (Hemming et al., 2011). This finding also proves that increasing the cluster size has a weaker effect on power than increasing the number of clusters, which is a characteristic also observed in the frequentist approach as illustrated in Equation (7).

Comparatively, when the effect size is 0.5, increasing the ICC leads to slight increases in cluster size for certain conditions, while for the majority, the cluster size remains unaffected. These conditions correspond to a threshold of 5 and cluster sizes of 30 or 60. Noteworthy is that we found similar patterns in the effect of the factors we have mentioned, regardless of whether the effect size is 0.5 or 0.8. The reader can use the Shiny app to verify that, for an effect size of 0.8, cluster sizes tend to be 6 and increase slightly

Table 2

Required number of individuals per cluster with a fraction of information equal to 1 and probability (η) of 0.8 for Hypotheses

Set 1

ICC = 0.025												ICC = 0.050				ICC = 0.100			
			Eff. Size = 0.2		Eff. Size = 0.5		Eff. Size = 0.2		Eff. Size = 0.5		Eff. Size = 0.2		Eff. Size = 0.5						
BF_{thresh}	n_2	Hypothesis	n_1	$P(BF > BF_{thresh})$	n_1	$P(BF > BF_{thresh})$	n_1	$P(BF > BF_{thresh})$	n_1	$P(BF > BF_{thresh})$	n_1	$P(BF > BF_{thresh})$	n_1	$P(BF > BF_{thresh})$					
1	30	H0	37	0.981	6	0.961	304	0.981	6	0.956	1000	0.971	6	0.950					
		H1		0.808		0.983		0.802		0.974		0.546		0.947					
1	60	H0	12	0.985	6	0.973	18	0.980	6	0.969	137	0.981	6	0.967					
		H1		0.803		1.000		0.816		1.000		0.800		0.998					
1	90	H0	8	0.984	6	0.979	9	0.981	6	0.978	16	0.981	6	0.975					
		H1		0.836		1.000		0.811		1.000		0.801		1.000					
3	30	H0	66	0.938	6	0.847	1000	0.933	6	0.834	1000	0.897	6	0.810					
		H1		0.806		0.953		0.692		0.929		0.390		0.880					
3	60	H0	17	0.942	6	0.907	29	0.938	6	0.894	1000	0.938	6	0.883					
		H1		0.804		0.999		0.803		0.998		0.703		0.994					
3	90	H0	10	0.939	6	0.927	14	0.943	6	0.924	40	0.943	6	0.914					
		H1		0.803		1.000		0.818		1.000		0.800		0.999					
5	30	H0	85	0.898	14	0.801	1000	0.871	22	0.806	1000	0.789	1000	0.789					
		H1		0.805		0.999		0.637		0.999		0.332		0.998					
5	60	H0	20	0.897	6	0.802	39	0.893	7	0.811	1000	0.874	10	0.806					
		H1		0.809		0.999		0.806		0.998		0.649		0.998					
5	90	H0	12	0.899	6	0.865	16	0.897	6	0.855	81	0.893	6	0.830					
		H1		0.823		1.000		0.805		1.000		0.800		0.999					

n_1 represents the cluster size and n_2 represents the number of clusters.

with the ICC under specific conditions. However, there are conditions in which the maximum cluster size is reached, particularly with high ICC ($\rho = 0.1$) and low number of clusters ($n_2 = 30$).

The relationship between the variables in Table 2 are similar to those described in Table 1. Larger cluster sizes are required when the effect size and the number of clusters decrease. There is a notable difference in the probability η according to the effect size: with a small effect size, the increase in ICC leads to a considerable increase in cluster size.

In general, larger cluster sizes were required with increasing the Bayes factor threshold and the fraction b . However, for conditions with a small effect size, the cluster size decreased or exhibited a different pattern of changes as the fraction b increased.

Determining the number of clusters for Hypotheses Set 2

Table 3 is different from the tables presented for Hypotheses Set 1. The consideration of hypotheses with only inequality constraints requires the simulation and test for one hypothesis given that, in this case, we are testing one hypothesis (i.e., H_1) against its complement. Moreover, as there is no equality constraint, the fraction of information b is not necessary for a sensitivity analysis. Regarding all the other factors, the interpretation of the table is similar to the interpretation for Table 1.

The table clearly indicates that, regardless of the Bayes factor threshold, ICC, effect size, and cluster size, the required number of clusters per treatment condition is 6. The cases that deviate from this tendency have the largest thresholds, the smallest cluster sizes, and the smallest effect sizes. This tendency means that, while the number of clusters may be larger in specific conditions, overall, the power criterion is easily met, with a number of clusters close to the minimum specified in the design of the simulation study.

Determining the cluster size for Hypotheses Set 2

Table 4 presents the cluster size as a function of the number of clusters, ICC, effect sizes, and Bayes factor thresholds. From the table can be inferred that, regardless of the ICC and threshold, the required cluster size is 8 when the effect size is 0.5. However, in the

Table 3

Required number of clusters per treatment condition with probability (η) of 0.8 for Hypotheses Set 2

ICC = 0.025												ICC = 0.050						ICC = 0.100					
			Eff. Size = 0.2			Eff. Size = 0.5			Eff. Size = 0.2			Eff. Size = 0.5			Eff. Size = 0.2			Eff. Size = 0.5					
$BF_{threshold}$	n_1	n_2	$P(BF > BF_{threshold})$	n_2	$P(BF > BF_{threshold})$	n_2	$P(BF > BF_{threshold})$	n_2	$P(BF > BF_{threshold})$	n_2	$P(BF > BF_{threshold})$	n_2	$P(BF > BF_{threshold})$	n_2	$P(BF > BF_{threshold})$	n_2	$P(BF > BF_{threshold})$						
1	30	6	0.969	6	1.000	6	0.959	6	1.000	6	0.942	6	1.000	6	1.000								
1	60	6	0.994	6	1.000	6	0.992	6	1.000	6	0.987	6	1.000	6	1.000								
1	90	6	0.999	6	1.000	6	0.998	6	1.000	6	0.996	6	1.000	6	1.000								
3	30	6	0.867	6	1.000	6	0.854	6	1.000	6	0.811	6	0.999	6	0.999								
3	60	6	0.967	6	1.000	6	0.957	6	1.000	6	0.936	6	1.000	6	1.000								
3	90	6	0.992	6	1.000	6	0.987	6	1.000	6	0.978	6	1.000	6	1.000								
5	30	7	0.830	6	1.000	8	0.818	6	0.999	12	0.806	6	0.998	6	0.998								
5	60	6	0.937	6	1.000	6	0.922	6	1.000	6	0.887	6	1.000	6	1.000								
5	90	6	0.983	6	1.000	6	0.975	6	1.000	6	0.959	6	1.000	6	1.000								
n_1 represents the cluster size and n_2 represents the number of clusters.																							

n_1 represents the cluster size and n_2 represents the number of clusters.

cases where the effect size is 0.2, larger cluster sizes are necessary when the number of clusters is low, the ICC increases, or the threshold increases.

The comparison of the results for Hypotheses Set 1 and Hypotheses Set 2 indicates that including hypotheses with equality constraints requires larger sample sizes. This outcome is obvious, since finding support for one specific value of the difference in group means ($H_0 : \mu_C - \mu_T = 0$) is more difficult than finding support for a range of values in the difference in group means ($H_A : \mu_C - \mu_T < 0$). Hence, H_1 has a better fit when the posterior distribution deviates from the specific value in H_0 .

Another difference in the results of the two hypothesis sets lies in that the required sample size for Hypotheses Set 2 hardly depends on the ICC, effect size, and Bayes factor threshold. The results for Hypotheses Set 1 are more consistent with the effects of the factors in the frequentist framework, which is expressed in equations (6) and (7) and proved in Moerbeek and Teerenstra (2016). On the other hand, the same relationship between the factors and the sample size is observed in specific conditions for Hypotheses Set 2, such as the smallest effect size, the largest thresholds, and the smallest fixed sample sizes.

Practical example

This section uses the example of the cluster-randomized trial carried out by Ausems et al. (2002), presented above. In this study, schools were assigned randomly to four treatment conditions to evaluate two interventions and their interaction. One of the variables that was measured is the attitude toward the disadvantages of smoking, which is a result of the sum of an 11-item scale with a 5-point Likert scale, ranging from 1=very negative to 5=very positive.

Suppose that a researcher wants to replicate the study of Ausems et al. (2002) but is only interested in the effect of the out-of-school condition versus the control. Attitude toward the disadvantages of smoking is the outcome variable for which a power analysis is to be performed. The pair of hypotheses to consider is

Table 4
Required number of individuals per cluster with probability (η) of 0.8 for Hypotheses Set 2

		ICC = 0.025				ICC = 0.050				ICC = 0.100			
		Eff. Size = 0.2		Eff. Size = 0.5		Eff. Size = 0.2		Eff. Size = 0.5		Eff. Size = 0.2		Eff. Size = 0.5	
BF_{thresh}	n_2	n_1	$P(BF > BF_{thresh})$	n_1	$P(BF > BF_{thresh})$	n_1	$P(BF > BF_{thresh})$	n_1	$P(BF > BF_{thresh})$	n_1	$P(BF > BF_{thresh})$	n_1	$P(BF > BF_{thresh})$
1	5	8	0.810	8	0.984	10	0.816	8	0.980	10	0.807	8	0.972
1	10	8	0.867	8	0.998	8	0.842	8	0.995	8	0.809	8	0.989
1	40	8	0.963	8	1.000	8	0.928	8	1.000	8	0.873	8	0.998
3	5	26	0.810	8	0.924	28	0.809	8	0.908	32	0.807	8	0.886
3	10	14	0.800	8	0.983	18	0.820	8	0.974	22	0.809	8	0.944
3	40	8	0.863	8	1.000	10	0.837	8	0.999	16	0.832	8	0.983
5	5	36	0.803	8	0.867	42	0.817	8	0.851	46	0.801	8	0.817
5	10	22	0.828	8	0.965	24	0.806	8	0.943	32	0.805	8	0.898
5	40	10	0.845	8	1.000	14	0.835	8	0.997	20	0.802	8	0.967

n_1 represents the cluster size and n_2 represents the number of clusters.

$$H_0 : \mu_{out} = \mu_{control}$$

$$H_1 : \mu_{out} > \mu_{control}.$$

The researcher performs sample size calculations. Following Moerbeek (2006), the between-cluster variance is equal to $\sigma_u^2 = 3.5$ and within-cluster variance is equal to $\sigma_e^2 = 45$, which means that the total variance is $\sigma^2 = 48.5$ and the ICC is $\rho = \frac{3.5}{48.5} = 0.0721$. The unstandardized effect size that the researcher is expecting to detect is 1.39, corresponding to the standardized effect size $\delta = 1.39/\sqrt{48.5} = 0.19$. The desired significance level is $\alpha = 0.05$ and statistical power $1 - \beta = 0.8$. If the cluster size is $n_1 = 30$ students per cluster, using the formula (6), the required total number of schools is

$$n_2 = 4 \frac{1 + (30 - 1)0.0721}{30} \left(\frac{1.96 + 0.84}{0.19} \right)^2 = 81.24714, \quad (12)$$

which is rounded up to 82.

However, the researcher is also open to using the Bayes factor as the method to test the hypotheses. For this reason, the researcher performs a sample size calculation to test the hypotheses with different threshold and probability values. Considering that the researcher wants to confirm the effects that have been studied before, the Bayes factor thresholds are 3 and 5 and the thresholds are 0.8 and 0.9.

The results in Table 5 show that the required number of clusters per treatment condition can be from 84 up to 252, increasing with the threshold (BF_{thresh}) and probability(η). The original study by Ausems et al. (2002) included 143 schools, which is higher than most of the required number of schools as listed in Table 5. The effect of increasing the fraction of information b on the required number of clusters may be increasing, decreasing, or constant, depending on the values of effect sizes and BF_{thresh} such as this study ($\delta = 0.19$).

Table 5

Required number of clusters for evaluation of smoking prevention programs with the Bayes factor when the cluster size equals 30 students.

η	BF_{thresh}	b	n_2	$P(BF_{01} > BF_{thresh} \mid H_0)$	$P(BF_{10} > BF_{thresh} \mid H_1)$
0.8	3	1	84	0.948	0.804
0.8	3	2	78	0.913	0.801
0.8	3	3	76	0.879	0.806
0.9	3	1	110	0.951	0.902
0.9	3	2	104	0.925	0.905
0.9	3	3	100	0.906	0.903
0.8	5	1	94	0.912	0.811
0.8	5	2	88	0.845	0.808
0.8	5	3	96	0.801	0.854
0.9	5	1	120	0.920	0.908
0.9	5	2	170	0.900	0.988
0.9	5	3	252	0.902	1.000

Note: η represents the probability that must be reached to meet the power criterion.

BF_{thresh} represents the Bayes factor threshold. In addition, n_2 stands for the required number of clusters to reach the power criterion.

Discussion

This paper presents an innovative method for determining the sample size for CRTs based on Bayesian power. The method is designed to determine the number of clusters or the cluster sizes when one of them is fixed. The Bayesian power criterion specifies that the sample size is determined so that it ensures the probability (η) of obtaining a Bayes factor larger than a Bayes factor threshold for either hypothesis under consideration.

To facilitate this approach, we have developed two functions, `SSD_crt_null` and

`SSD_crt_inform`, that are implemented and freely available in an R package called `SSD_Bayes_ML`. The first function determines the required sample size for the evaluation of Hypothesis Set 1, where $H_0 : \mu_C = \mu_T$ and $H_1 : \mu_C < \mu_T$ are evaluated. The second function determines the required sample size to evaluate Hypotheses Set 2, which is $H_1 : \mu_C < \mu_T$ against its complement $H_2 : \mu_C > \mu_T$. In addition, the results from the simulations can be consulted and explored in the Shiny app Sample Size Determination for Cluster Randomised Trials with Bayes Factor (<https://utrecht-university.shinyapps.io/BayesSamplSizeDet-CRT/>).

The results showed the effect of ICC, effect size, and fixed sample size on the determination of the required sample size. Larger sample sizes are required when the ICC increases, the fixed sample size is small, or the effect size is small. These results align with the effect of the same factors on the sample size determination in the frequentist approach (Moerbeek & Teerenstra, 2016; Raudenbush, 1997b). The simulation showed the trade-off between the number of clusters and cluster sizes; in addition and as expected (Hemming et al., 2011), the required power level η could not always be achieved for a limited number of clusters, even when the cluster size was as large as 1,000.

The effects of the Bayes factor threshold and fraction of information b on the required sample size indicate that larger sample sizes were required for larger values of thresholds and fraction of information. However, important to note is that the tendency of the effect of fraction of information on sample size determination also depends on the values of the effect size and ICC. In addition, this tendency only happens when a hypothesis with equality constraint is evaluated. Alternatively, our findings for Hypothesis Set 2 suggest that the required sample sizes tended to be small and varied little, regardless of the factor levels used in the simulation study. This tendency lends additional support to questioning the testing of the null hypothesis (for further arguments against the use of NHST, see: David R. Anderson et al., 2000; De Schoot et al., 2011; Gu et al., 2014; Hoijtink, Mulder, et al., 2019; Klugkist et al., 2011).

The simulations exhibited the differences in reaching the power criterion for different hypotheses. The power criterion was easily met when evaluating Hypotheses Set 2 for all conditions. Moreover, the required sample size had a small range of values and varied little, especially for cases with medium and large effect sizes. In comparison, when evaluating Hypotheses Set 1, the required sample size varied considerably, and in cases with small effect sizes and a small number of clusters the power criterion was not met, even after reaching 1,000 individuals per cluster.

The method and the software presented in this paper implement the approximated adjusted fractional Bayes factor (AAFBF), which is only one type of Bayes factor. While one of the advantages of the Bayes factor is the incorporation of prior information, for this type of Bayes factor the user only has to indicate the fraction of information from data used to specify the prior distribution. However, the AAFBF is sensible to different fractions of the sample in the case of equality-constrained hypotheses. Additionally, the software presented in this paper is specifically tailored for determining the sample sizes in a parallel cluster-randomized trial with only two treatment conditions and evaluating two hypotheses. Additional elements in CRT design that also influence the determination of sample size include unequal cluster sizes, uncertainty surrounding intraclass correlation, and non-inferiority and equivalence designs (Rutterford et al., 2015). For NHST, the loss of efficiency due to variation in cluster sizes has been shown to rarely exceed 10%, meaning 11% more clusters should be added to compensate (Van Breukelen et al., 2007). This paper is restricted to equal allocation ratios, which can be shown to be optimal when costs and variances do not vary across treatment conditions (Schouten, 1999).

In general, the methods for sample size determination also requires an educated guess of the ICC. This value can be obtained from the literature or expert knowledge. For instance, Table 11.1 in Moerbeek and Teerenstra (2016) shows a summary of papers that report estimates of ICCs in CRTs in various fields of science. One may consider a sensitivity analysis and determine the sample size for a range of plausible values of the ICC

to study the degree to which sample size depends on the ICC. Future research may focus on sample size determination variances varying across treatment conditions.

Another consideration when utilizing the provided functions is computational cost, since our method to determine the required sample relies on simulations. In general, the minimum running time is approximately 5 minutes, while the largest running time is 35 hours. These results were obtained with a 16-core GPU of 250GB of RAM. To improve efficiency and reduce computation time, the functions in the package employ a binary search algorithm to find the required sample size. However, it is important to highlight that the combinations with the largest running time always corresponded to the evaluation of Hypothesis Set 1. Most likely, in the near future, this limitation of computational cost will be solved with the advance of technology.

Considering the growing popularity of Bayes factors for hypothesis testing in psychology (Heck et al., 2022), our method for sample size determination in CRTs is an important advance in research. We have previously discussed the disadvantages of null hypothesis significance testing (NHST), and how the Bayes factors provide an alternative approach to hypothesis testing that can avoid these disadvantages. One of the drawbacks of NHST is that, in practice, the decision to reject or maintain the null hypothesis relies on an arbitrary level of significance. To avert the same misinterpretation of the thresholds in our method, important to note is that the Bayes factor is the quantification of the evidence of the hypotheses under consideration and it may take values from 0 to infinity. Although the thresholds used in this paper are often seen in practice, we encourage the researchers to choose the thresholds based on the aimed-for degree of support.

This paper introduced a method for sample size determination in cluster-randomized trials (CTRs) tailored for hypothesis testing using Bayes factors. Moreover, we provided the practical implementation of the method through functions in the R package `SSD_Bayes_ML`. To our knowledge, this was the first contribution to Bayesian sample size determination for CRTs. This paper is part of a larger four-year

project; in the years to come, we aim to focus on trials with more than two treatment conditions (thereby extending the paper by Fu (2022) to the multilevel setting) and cluster-randomized crossover. We also consider alternative measures for the evaluation of hypotheses, such as the GORICA (Altinisik et al., 2021; Vanbrabant et al., 2020).

References

- Altinisik, Y., Van Lissa, C. J., Hoijsink, H., Oldehinkel, A. J., & Kuiper, R. M. (2021). Evaluation of inequality constrained hypotheses using a generalization of the AIC. *Psychological Methods*, 26(5), 599–621. <https://doi.org/10.1037/met0000406>
- Ausems, M., Mesters, I., Van Breukelen, G., & De Vries, H. (2002). Short-Term Effects of a Randomized Computer-Based Out-of-School Smoking Prevention Trial Aimed at Elementary Schoolchildren. *Preventive Medicine*, 34(6), 581–589. <https://doi.org/10.1006/pmed.2002.1021>
- Balluerka, N., Gómez, J., & Hidalgo, D. (2005). The Controversy over Null Hypothesis Significance Testing Revisited. *Methodology*, 1(2), 55–70. <https://doi.org/10.1027/1614-1881.1.2.55>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1). <https://doi.org/10.18637/jss.v067.i01>
- Binary search algorithm. (2024). *Wikipedia*
Page Version ID: 1218941965.
- Campbell, M. J., & Walters, S. J. (2014, April). *How to Design, Analyse and Report Cluster Randomised Trials in Medicine and Health Related Research* (1st ed.). Wiley. <https://doi.org/10.1002/9781118763452>
- David R. Anderson, Anderson, D. E., David R. Anderson, Kenneth P. Burnham, Burnham, K. P., & Thompson, W. L. (2000). Null hypothesis testing: Problems, prevalence, and an alternative. *Journal of Wildlife Management*, 64(4), 912–923. <https://doi.org/10.2307/3803199>
MAG ID: 1991445071 S2ID: 42ed4a082b4ef252dabc67a4af43a74fffd8d543.
- De Schoot, R. V., Hoijsink, H., & Jan-Willem, R. (2011). Moving Beyond Traditional Null Hypothesis Testing: Evaluating Expectations Directly. *Frontiers in Psychology*, 2. <https://doi.org/10.3389/fpsyg.2011.00024>

- Donner, A., & Klar, N. (2010). *Design and analysis of cluster randomization trials in health research*. Wiley.
- Eldridge, S., & Kerry, S. (2012, January). *A Practical Guide to Cluster Randomised Trials in Health Services Research* (1st ed.). Wiley. <https://doi.org/10.1002/9781119966241>
- Fu, Q. (2022, March). *Sample Size Determination for Bayesian Informative Hypothesis Testing*: [Doctoral dissertation, Utrecht University]. <https://doi.org/10.33540/1221>
- Fu, Q., Hoijtink, H., & Moerbeek, M. (2021). Sample-size determination for the Bayesian t test and Welch's test using the approximate adjusted fractional Bayes factor. *Behavior Research Methods*, 53(1), 139–152. <https://doi.org/10.3758/s13428-020-01408-1>
- Gelfand, A. E., & Wang, F. (2002). A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models. *Statistical Science*, 17(2), 193–208. <https://doi.org/10.1214/ss/1030550861>
- Goldstein, H. (2011). *Multilevel statistical models* (4th ed). Wiley.
- Gu, X., Mulder, J., Deković, M., & Hoijtink, H. (2014). Bayesian evaluation of inequality constrained hypotheses. *Psychological Methods*, 19(4), 511–527. <https://doi.org/10.1037/met0000017>
- Gu, X., Mulder, J., & Hoijtink, H. (2018). Approximated adjusted fractional Bayes factors: A general method for testing informative hypotheses. *British Journal of Mathematical and Statistical Psychology*, 71(2), 229–261. <https://doi.org/10.1111/bmsp.12110>
- Hayes, R. J., & Moulton, L. H. (2009). *Cluster randomised trials*. CRC Press. OCLC: 244246621.
- Heck, D. W., Boehm, U., Böing-Messing, F., Bürkner, P.-C., Derks, K., Dienes, Z., Fu, Q., Gu, X., Karimova, D., Kiers, H. A. L., Klugkist, I., Kuiper, R. M., Lee, M. D., Leenders, R., Leplaa, H. J., Linde, M., Ly, A., Meijerink-Bosman, M., Moerbeek, M., ... Hoijtink, H. (2022). A review of applications of the Bayes factor

- in psychological research. *Psychological Methods*.
<https://doi.org/10.1037/met0000454>
- Hemming, K., Girling, A. J., Sitch, A. J., Marsh, J., & Lilford, R. J. (2011). Sample size calculations for cluster randomised controlled trials with a fixed number of clusters. *BMC Medical Research Methodology*, 11(1), 102.
<https://doi.org/10.1186/1471-2288-11-102>
- Hemming, K., & Kasza, J. (n.d.). The Shiny CRT Calculator: Power and Sample size for Cluster Randomised Trials.
- Hoijtink, H. (2012). *Informative hypotheses: Theory and practice for behavioral and social scientists*. CRC.
- Hoijtink, H., Gu, X., & Mulder, J. (2019). Bayesian evaluation of informative hypotheses for multiple populations. *British Journal of Mathematical and Statistical Psychology*, 72(2), 219–243. <https://doi.org/10.1111/bmsp.12145>
- Hoijtink, H., Mulder, J., Van Lissa, C., & Gu, X. (2019). A tutorial on testing hypotheses using the Bayes factor. *Psychological Methods*, 24(5), 539–556.
<https://doi.org/10.1037/met0000201>
- Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2017, September). *Multilevel Analysis: Techniques and Applications* (3rd ed.). Routledge.
<https://doi.org/10.4324/9781315650982>
- Jeffreys, H. (1983). *Theory of probability* (3rd ed). Clarendon Press ; Oxford University Press.
- Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, 90(430), 773–795. <https://doi.org/10.1080/01621459.1995.10476572>
- Klugkist, I., & Hoijtink, H. (2007). The Bayes factor for inequality and about equality constrained models. *Computational Statistics & Data Analysis*, 51(12), 6367–6379.
<https://doi.org/10.1016/j.csda.2007.01.024>

- Klugkist, I., Laudy, O., & Hoijtink, H. (2005). Inequality Constrained Analysis of Variance: A Bayesian Approach. *Psychological Methods*, 10(4), 477–493.
<https://doi.org/10.1037/1082-989X.10.4.477>
- Klugkist, I., Van Wesel, F., & Bullens, J. (2011). Do we know what we test and do we test what we want to know? *International Journal of Behavioral Development*, 35(6), 550–560. <https://doi.org/10.1177/0165025411425873>
- Lazega, E., & Snijders, T. A. B. (Eds.). (2016). *Multilevel network analysis for the social sciences: Theory, methods and applications*. Springer.
- Moerbeek, M. (2005). Randomization of Clusters Versus Randomization of Persons Within Clusters: Which Is Preferable? *The American Statistician*, 59(1), 72–78.
<https://doi.org/10.1198/000313005X20727>
- Moerbeek, M. (2006). Power and money in cluster randomized trials: When is it worth measuring a covariate? *Statistics in Medicine*, 25(15), 2607–2617.
<https://doi.org/10.1002/sim.2297>
- Moerbeek, M., & Teerenstra, S. (2016). *Power analysis of trials with multilevel data*. CRC Press, Taylor & Francis.
- Moerbeek, M., Van Breukelen, G. J. P., & Berger, M. P. F. (2000). Design Issues for Experiments in Multilevel Populations. *Journal of Educational and Behavioral Statistics*, 25(3), 271. <https://doi.org/10.2307/1165206>
- Mulder, J. (2014). Prior adjusted default Bayes factors for testing (in)equality constrained hypotheses. *Computational Statistics & Data Analysis*, 71, 448–463.
<https://doi.org/10.1016/j.csda.2013.07.017>
- Murray, D. M. (1998). *Design and analysis of group-randomized trials*. Oxford university press.
- Oakes, M. W. (1987). *Statistical inference: A commentary for the social and behavioural sciences* (Reprint). Wiley.

- O'Hagan, A. (1995). Fractional Bayes Factors for Model Comparison. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 99–118.
<https://doi.org/10.1111/j.2517-6161.1995.tb02017.x>
- Raudenbush, S. W. (1997a). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, 2(2), 173–185.
<https://doi.org/10.1037/1082-989X.2.2.173>
- Raudenbush, S. W. (1997b). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods*, 2(2), 173–185.
<https://doi.org/10.1037/1082-989X.2.2.173>
- Raudenbush, S. W., & Bryk, A. S. (2010). *Hierarchical linear models: Applications and data analysis methods* (2. ed., [Nachdr.]). Sage Publ.
- Rutterford, C., Copas, A., & Eldridge, S. (2015). Methods for sample size determination in cluster randomized trials. *International Journal of Epidemiology*, 44(3), 1051–1067.
<https://doi.org/10.1093/ije/dyv113>
- Schouten, H. J. A. (1999). Sample size formula with a continuous outcome for unequal group sizes and unequal variances. *Statistics in Medicine*, 18(1), 87–91. [https://doi.org/10.1002/\(SICI\)1097-0258\(19990115\)18:1<87::AID-SIM958>3.0.CO;2-K](https://doi.org/10.1002/(SICI)1097-0258(19990115)18:1<87::AID-SIM958>3.0.CO;2-K)
- Van Breukelen, G. J. P., Candel, M. J. J. M., & Berger, M. P. F. (2007). Relative efficiency of unequal *versus* equal cluster sizes in cluster randomized and multicentre trials. *Statistics in Medicine*, 26(13), 2589–2603. <https://doi.org/10.1002/sim.2740>
- Vanbrabant, L., Van Loey, N., & Kuiper, R. M. (2020). Evaluating a theory-based hypothesis against its complement using an AIC-type information criterion with an application to facial burn injury. *Psychological Methods*, 25(2), 129–142.
<https://doi.org/10.1037/met0000238>
- Wilson, K. J. (2022, August). Bayesian design and analysis of two-arm cluster randomised trials using assurance.

Appendix A

Calculation of Approximated Adjusted Fractional Bayes Factor

According to Fu et al. (2021), Gu et al. (2018), and Hoijtink, Gu, and Mulder (2019), the adjusted fractional prior distribution is given by

$$\pi_u^*(\theta|X^b) = N \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \frac{1}{b} \frac{\sigma_C^2}{N_{eff}} & 0 \\ 0 & \frac{1}{b} \frac{\sigma_T^2}{N_{eff}} \end{bmatrix} \right) \quad (\text{A1})$$

Where σ_C^2 represents the variance of the control condition and σ_T^2 represents the variance of the treatment condition. One can observe that fraction b affects the variance of the prior distribution, resulting in a larger variance of the distribution with small fractions and a lower variance with large fractions. As explained above, due to the dependence of the observations, the effective sample size is used (N_{eff}), which according to Hox et al. (2017) is defined as

$$N_{eff} = \frac{n_2 n_1}{1 + (n_1 - 1)\rho}. \quad (\text{A2})$$

The normal approximation of the posterior distribution is

$$\pi_u(\theta|X) = N \left(\begin{bmatrix} \mu_C \\ \mu_T \end{bmatrix}, \begin{bmatrix} \frac{\sigma_C^2}{N_{eff}} & 0 \\ 0 & \frac{\sigma_T^2}{N_{eff}} \end{bmatrix} \right) \quad (\text{A3})$$

where μ_C denotes the mean of control condition and μ_T denotes the maximum likelihood estimate of the mean of treatment condition.

Choosing b

The fraction of information b has numerous definitions depending on the type of Bayes factor that is used and the researcher's aim. In this paper, we used the Approximated Adjusted Fractional Bayes Factor, for which the minimal training sample is defined as

$$b = \frac{J}{N_{eff}}. \quad (\text{A4})$$

where J is the number of restrictions in the hypotheses under consideration. However, there are other definitions of b ; researchers interested in this definitions can refer to Gu et al. (2018).

Appendix B

Binary search algorithm

The binary search algorithm reduces the number of iterations and hence the computation time in the algorithm proposed in section Methodology for Sample Size Determination.

The binary search takes values (N_{mid}) between N_{min} and N_{max} until the power criteria of $P(BF_{01} > BF_{thresh}|H_0) > \eta$ and $P(BF_{10} > BF_{thresh}|H_1) > \eta$ are met. As a result, the time complexity is reduced to $O(\log N)$ (“Binary Search Algorithm,” 2024), which means that, in the worst scenario, the algorithm makes $\log_2(1000 - 5) + 1 = 10$ evaluations of the power criterion.

1. N_{max} is specified by the user as an argument of the R function, while N_{min} is defined inside the R function.
2. The midpoint is calculated as $N_{mid} = \frac{N_{min} + N_{max}}{2}$.
3. If the power criterion in Equation (11) is satisfied, the maximum is updated to $N_{max} = N_{mid}$ and N_{mid} is calculated again. Otherwise, the minimum is set to $N_{min} = N_{mid}$, and N_{mid} is recalculated. The result of this recalculation means that in the first case, the sample size decreases and in the second case, the sample size increases.
4. The algorithm terminates under two conditions. For cluster size determination, the algorithm stops when $N_{mid} = N_{min} + 1$ and $N_{min} - 1$ fails to satisfy the power criterion. In the case of determining the number of clusters, it stops when $N_{mid} = N_{min} + 2$ and $N_{min} - 2$ fails to meet the power criterion. In either case, upon termination $N = N_{mid}$.