

Methodology and Statistics for the Behavioural, Biomedical, and Social  
Sciences  
Utrecht University, the Netherlands

MSc Thesis Fayette Klaassen (4104803)  
TITLE: The Power of Informative Hypotheses  
May 2015

Supervisors:  
Prof. Dr. Herbert Hoijtink  
MSc Xin Gu

Preferred journal of publication: Psychological Methods  
Word count: 8605

# The Power of Informative Hypotheses

Fayette Klaassen\*, Xin Gu, and Herbert Hoijtink

*Department of Methodology and Statistics, Utrecht University*

## Abstract

Researchers can express their expectations regarding the ordering of group means in one or more simple order constrained hypotheses, for example  $H_i : \mu_1 > \mu_2 > \mu_3$ ,  $H_c : \text{not } H_i$ , and  $H_{i'} : \mu_3 > \mu_2 > \mu_1$ . They can compare these hypotheses by means of a Bayes factor. This article determines the required group sample size for the evaluation of  $H_i$  with  $H_c$  or  $H_{i'}$  by means of a Bayes factor. Three approaches have been developed for sample size determination that depend on different decision criteria. The first approach makes a dichotomous decision for one of the hypotheses considered. The second approach makes a trichotomous decision for either one of the hypotheses considered or decides that not enough support was found in the data for either hypothesis. The third approach describes the support found in the data for each hypothesis considered. Simulations were performed to determine the sample size such that error probabilities are acceptably low or expected evidence is acceptably strong. Comparisons between the use of  $H_c$  and  $H_{i'}$  are made. First, the required sample size decreases if  $H_i$  is compared with  $H_{i'}$  instead of  $H_c$ . Thus, specifying what orderings of means are expected or are of interest decreases the required sample size. Second, the required sample sizes differ over the three approaches. The choice for an approach is, amongst others, dependent on the type of conclusion a researcher wants to obtain. A decision tree is provided to guide researchers to the appropriate approach. Applied researchers can use the decision tree and the tables presented to determine the required sample size for their research or use R code and associated manual provided in this paper.

**Keywords:** ANOVA; Bayes factor; informative hypotheses; power; sample size.

## 1 Introduction

In a classical ANOVA, researchers compare a null hypothesis  $H_0$  with an alternative hypothesis  $H_1$ :

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_K, \quad (1)$$

$$H_1 : \mu_1 \neq \mu_2 \neq \dots \neq \mu_K, \quad (2)$$

---

\*Corresponding author. Email: klaassen.fayette@gmail.com Address: Department of Methodology and Statistics, Utrecht University, P.O. Box 80140, 3508 TC Utrecht, The Netherlands. FK wrote the paper, R code, and executed the simulations. XG and HH conceptualized the project, discussed steps, and provided feedback on writing.

where  $\mu_k$  indicates the mean in group  $k$  for  $k = 1, 2, \dots, K$ , and  $K$  indicates the number of groups. A researcher might not be interested in  $H_0$  and  $H_1$  (cf. Cohen, 1994; Rozeboom, 1997), but in testing a theory concerning the order of group means (Van de Schoot, Hoijtink, and Romeijn, 2011; Gu, Mulder, Deković, and Hoijtink, 2014). He can express his expectations with regard to the ordering of these means in an informative hypothesis  $H_i$ . In this paper, only simple order constrained hypotheses (Kuiper & Hoijtink, 2010) are considered:

$$H_i : \mu_1 > \mu_2 > \dots > \mu_K. \quad (3)$$

A researcher can compare  $H_i$  with another order constrained hypothesis  $H_{i'}$ , for example:

$$H_{i'} : \mu_2 > \mu_1 > \dots > \mu_K, \quad (4)$$

or with  $H_c$ , the complement of  $H_i$ :

$$H_c : \text{not } H_i. \quad (5)$$

In a Bayesian framework, these informative hypotheses can be evaluated with a Bayes factor (Kass and Raftery, 1995; Hoijtink, 2012, p. 50-51).  $BF_{ic}$  expresses the support in the data for  $H_i$  relative to  $H_c$ . For example, when  $BF_{ic} = 5$ , the support in the data for  $H_i$  is 5 times stronger than for  $H_c$ . When  $BF_{ic} = 0.1$ , the support for  $H_c$  is 10 times stronger than for  $H_i$ . This paper will determine the sample sizes needed for a ‘powerful’ evaluation of  $H_i$  and  $H_c$  or  $H_i$  and  $H_{i'}$  using a Bayes factor. In the sequel it will be elaborated what meaning is given to powerful. First, an overview of what is the state of the art for sample size determination is given, and it is explained how our research relates to this.

Cohen (1988) has developed power analyses for, amongst others, ANOVA models. These analyses describe the interdependence of power, Type I error, effect size, and sample size. However, Cohen’s power analysis only applies to the comparison of a null hypothesis with a one- or two-sided alternative by means of the p-value. We are interested in the comparison of  $H_i$  with  $H_c$  or with  $H_{i'}$  by means of a Bayes factor.

Adcock (1997) gives an overview of sample size determination methods that have been developed for the evaluation of two hypotheses by means of a Bayes factor. Amongst others, he discusses the method of Weiss (1997). Weiss (1997) evaluates  $H_0 : \mu = 0$  with  $H_1 : \mu \neq 0$ , or the one-sided variant of  $H_1$  by means of the logarithm of  $BF_{01}$ . He does not determine sample size, but evaluates the effect of a decision criterion on error probabilities, for four different sample sizes. He describes five decision criteria when using  $\log BF_{01}$ , and applies two of these approaches. In the first approach a critical  $\log BF_{01}$  is determined for each of the sample sizes, such that the Type I error probability is .05. The second approach uses a critical  $\log BF_{01} = 0$ , where the logarithms of Bayes factors smaller or larger than zero result in a decision for  $H_1$  and  $H_0$  respectively. For each sample size, the corresponding power and Type I error are determined (Weiss, 1997).

Other decision criteria for sample size determination, also based on the comparison of

$H_0 : \mu = 0$  with  $H_1 : \mu \neq 0$  have been developed. De Santis (2004, 2007) describes a decision criterion where Bayes factors are only considered decisive if they are smaller than  $\frac{1}{3}$  or larger than 3, respectively. The sample size is determined such that  $P(BF_{01} > 3|H_0)$  and  $P(BF_{01} < \frac{1}{3}|H_1)$  are both larger than a pre-specified value. Reyes and Ghosh (2013) consider not only hypotheses regarding one mean, but also the difference between two means. One of their methods determines a critical Bayes factor such that the average error probability is minimized. The sample size is then determined such that this minimized error probability is smaller than a specified cut-off value. Each of these methods of sample size determination use the Bayes factor as a key decision criterion. However, none consider the comparison of  $H_i$  with  $H_c$  or with  $H_{i'}$ .

Sample size determination for the evaluation of  $H_0$  with  $H_i$  using  $BF_{i0}$  is considered by Klugkist, Post, Haarhuis, and van Wesel (2014). The decision criterion used is that Bayes factors larger and smaller than 1 result in conclusions in favour of  $H_i$  and  $H_0$  respectively. Using this decision criterion, the sample size is determined for various effect sizes, such that the traditional Type I error probability is below .05, and the power is above .80 (Klugkist et al., 2014). Although this article uses order constrained hypotheses, no elaboration is made on the sample sizes required for the evaluation of  $H_i$  with  $H_c$  or with  $H_{i'}$ .

Recently, Vanbrabant, van de Schoot, and Rosseel (2015) have developed sample size tables for the evaluation of amongst others simple order constrained hypotheses using the  $F$ -bar statistic (Silvapulle & Sen, 2004). They contrast  $H_i$  both with  $H_1$  to check for model misfit, and with  $H_0$  to check for an existing effect by using p-values. This method shows the potential of the use of informative hypotheses, by comparing their sample sizes with those following from sample size determination methods that do not incorporate order constraints. However, this method does not evaluate  $H_i$  with another informative hypothesis, and does not use the Bayes factor.

All existing methods for sample size determination do not allow for the evaluation of two order constrained hypotheses by means of a Bayes factor. Hoijtink (2012, p. 115–118) gives indications of appropriate sample sizes in this situation. We will elaborate on this research by developing sample size tables for evaluation of  $H_i$  and  $H_c$  or  $H_i$  and  $H_{i'}$ . We will develop three decision criteria for the Bayes factor, that incorporate aspects of the previously mentioned approaches. Two approaches are based on controlling error probabilities, whereas the third approach uses a median Bayes factor of a required size. Guidelines for choosing between the three approaches and using the sample size tables are provided.

The remainder of this paper is organized as follows. First, an introduction of the Bayes factor is provided. Then the three decision criteria upon which sample size determination will be based, are discussed. Subsequently, the simulation procedure by which the sample sizes are determined is explained step by step. Next, the sample size tables are presented and discussed, and guidelines for using these tables are provided. The paper is concluded with a short discussion.

## 2 Bayes factor

The Bayes factor is a tool for Bayesian hypothesis testing. Bayes factors can be computed for every pair of hypotheses, and can be used to quantify the evidence in favour of one of these hypotheses. Bayes factors penalize the fit with the complexity of the hypotheses under consideration, where the fit describes how well the data support a hypothesis, and the complexity describes how specific a hypothesis is. The Akaike Information Criterion (AIC) for example, is based on a similar principle. The AIC penalizes the maximum value of the likelihood, which is a measure of fit, by the number of parameters, which is a measure of complexity (Akaike, 1973).  $BF_{i1}$  can be expressed as a ratio of the fit  $f_i$  and the complexity  $c_i$  of  $H_i$  and expresses the support in the data for  $H_i$  relative to  $H_1$  (Hojtink, 2012, p. 51–52):

$$BF_{i1} = \frac{f_i}{c_i}. \quad (6)$$

Using  $BF_{i1}$  and  $BF_{c1}$  or  $BF_{i'1}$ , Bayes factors can be obtained that express the support in the data for  $H_i$  relative to  $H_c$  or  $H_{i'}$ :

$$BF_{ic} = \frac{BF_{i1}}{BF_{c1}} = \frac{f_i}{c_i} / \frac{1 - f_i}{1 - c_i}, \quad (7)$$

$$BF_{ii'} = \frac{BF_{i1}}{BF_{i'1}} = \frac{f_i}{c_i} / \frac{f_{i'}}{c_{i'}}. \quad (8)$$

In order to compute the fit and complexity of a hypothesis, the density of the data, and the prior and posterior distributions are needed. For an ANOVA model, the density of the data is:

$$f(\mathbf{y}|\boldsymbol{\mu}, \sigma^2) = \prod_{k=1}^K \prod_{s=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp -\frac{1}{2} \frac{(y_{ks} - \mu_k)^2}{\sigma^2}, \quad (9)$$

where  $\mathbf{y} = [y_{11}, \dots, y_{1N}, \dots, y_{K1}, \dots, y_{KN}]$ ,  $\boldsymbol{\mu} = [\mu_1, \dots, \mu_K]$ ,  $\sigma^2$  indicates the within group variance and is equal for each group,  $k = 1, 2, \dots, K$  indicates a group, and  $s = 1, 2, \dots, N$  indicates a person in group  $k$ . The sample size, denoted by  $N$ , is equal for each group.

Based on Gu et al. (2014) a normal prior distribution is used for the parameters in the hypothesis, that is, the group means:

$$h(\boldsymbol{\mu}) = h(\mu_1) \cdot \dots \cdot h(\mu_K), \quad (10)$$

with

$$h(\mu_k) = \mathcal{N}(0, \infty),$$

for  $k = 1, \dots, K$ , in which the prior means are zero and the prior variances approach infinity.

This prior ensures two things. First, the influence of the prior on the posterior is

so small, that the posterior depends fully on the data. Secondly, we can use a normal approximation of the posterior distribution for the parameters used in the hypotheses, that is, the group means:

$$g(\boldsymbol{\mu}|\mathbf{y}) = g(\mu_1) \cdots g(\mu_K), \quad (11)$$

with

$$g(\mu_k|\mathbf{y}) = \mathcal{N}(\hat{\mu}_k, \hat{\tau}_k^2),$$

for  $k = 1, 2, \dots, K$ , in which  $\hat{\mu}_k$  is the estimate of the mean in group  $k$ , and  $\hat{\tau}_k^2$  is the squared standard error of the mean in group  $k$ , where

$$\hat{\mu}_k = \frac{1}{N} \sum_{s=1}^N y_{ks}, \quad (12)$$

$$\hat{\tau}_k^2 = \frac{\sum_{s=1}^N (y_{ks} - \hat{\mu}_k)^2}{N \cdot (N - 1)}. \quad (13)$$

The complexity and fit of a hypothesis are based on the prior and posterior distribution. The complexity of  $H_i$ ,  $c_i$ , describes how specific  $H_i$  is. It is the proportion of the prior distribution in agreement with  $H_i$  (Hoijtink, 2012, p. 60):

$$c_i = \int_{\boldsymbol{\mu} \in H_i} h(\boldsymbol{\mu}) d\boldsymbol{\mu}. \quad (14)$$

From Equation 14 it follows that  $c_i = 1/K!$ , and that  $c_c = 1 - c_i$  (Hoijtink, 2012, p. 60). Note that since  $H_c$  is the complement of  $H_i$ ,  $c_i + c_c = 1$ .

The fit of  $H_i$ ,  $f_i$ , describes how well the data support  $H_i$ . It is the proportion of the posterior distribution in agreement with  $H_i$  (Hoijtink, 2012, p. 59):

$$\begin{aligned} f_i &= \int_{\boldsymbol{\mu} \in H_i} g(\boldsymbol{\mu}|\mathbf{y}) d\boldsymbol{\mu} \\ &\approx \sum_{t=1}^T I_{\boldsymbol{\mu}_t \in H_i} / T, \end{aligned} \quad (15)$$

where  $\boldsymbol{\mu}_t$  is sampled from  $g(\boldsymbol{\mu}|\mathbf{y})$ ,  $I_{\boldsymbol{\mu}_t \in H_i}$  is 1 if  $\boldsymbol{\mu}_t$  is in agreement with  $H_i$ , and 0 otherwise, and  $T$  is the number of posterior samples. Again, since  $H_c$  is the complement of  $H_i$ , it follows that  $f_c = 1 - f_i$ . Using the complexity and fit, Bayes factors can be computed. The interpretation of Bayes factors is in terms of the relative support for each of a pair of hypotheses. For example, if  $BF_{ic} = 5$ , the support in the data for  $H_i$  is 5 times stronger than for  $H_c$ .

This paper develops three approaches that use the Bayes factor as a decision criterion to evaluate  $H_i$ ,  $H_{i'}$ , and  $H_c$ . All approaches make use of the sampling distributions of the Bayes factors under  $H_i$  and  $H_c$ , or under  $H_i$  and  $H_{i'}$ . Approach 1, like in Klugkist et al. (2014) and Weiss (1997), chooses  $H_i$  if  $BF_{ic} > 1$  or  $BF_{ii'} > 1$ , and chooses  $H_c$  if  $BF_{ic} < 1$  or  $H_{i'}$  if  $BF_{ii'} < 1$ . Sample sizes will be determined such that error

probabilities are acceptably low. Approach 2, like in De Santis (2004, 2007), chooses  $H_i$  if  $BF_{ic} > 3$  or  $BF_{ii'} > 3$ , and chooses  $H_c$  if  $BF_{ic} < \frac{1}{3}$  or  $H_{i'}$  if  $BF_{ii'} < \frac{1}{3}$ . No decision is made if Bayes factors are between  $\frac{1}{3}$  and 3. Again, sample sizes will be determined such that error probabilities are acceptably low. These error probabilities will be introduced in Section 3. In Approach 3, the Bayes factor is not used to make a decision, but to express support for  $H_i$  and  $H_c$  or  $H_{i'}$  based on the data. Sample sizes will be determined such that reasonably high Bayes factors can be expected, for example, 3, 10, or 20.

### 3 Decision criteria, error probabilities, and median Bayes factor

The sample size needed for the evaluation of  $H_i$  versus  $H_{i'}$  or versus  $H_c$  can be determined such that error probabilities are acceptably low, or the median Bayes factor under the true hypothesis expresses acceptably strong support. This section will first explain how sampling distributions of Bayes factors are obtained. Second, each approach is explained in more detail, by precisely defining error probabilities and the median Bayes factor required. Finally, it will be described what is meant by acceptably low error probabilities and strong support. Throughout this section, the comparison of  $H_i$  and  $H_c$  using  $BF_{ic}$  is discussed. The discussion is analogous for  $H_i$  and  $H_{i'}$ , where all comments and notations regarding  $H_c$  can be replaced with corresponding ones regarding  $H_{i'}$ .

All approaches in this paper make use of the sampling distributions of Bayes factors. Amongst others, the effect sizes under  $H_i$  and under  $H_c$  need to be defined to obtain the sampling distributions. In this paper, Cohen's  $d$ , the standardized difference between two means, is used as a measure of effect size (Cohen, 1988, p. 276). The effect size  $d_{H_i}$  under  $H_i$  is the standardized difference between the largest and the smallest mean under  $H_i$ .

$$d_{H_i} = \frac{\mu_K - \mu_1}{\sigma}, \quad (16)$$

where  $\mu_K$  is the largest mean, and  $\mu_1$  is the smallest mean under  $H_i$ . The effect size  $d_{H_c}$  under  $H_c$  is the standardized difference between the largest and the smallest population mean under  $H_c$ . For example, Figure 1a displays hypothetical sampling distributions of  $BF_{ic}$  under  $H_i$  and under  $H_c$ , given  $N = 50$ ,  $d_{H_i} = .2$ , and  $d_{H_c} = .2$ . These distributions represent the values of the Bayes factors observed if we repeatedly sample from populations under  $H_i$  and  $H_c$ . The procedure to obtain sampling distributions will be explained in full detail in Section 4.4.

#### 3.1 Approach 1

The decision criterion used in Approach 1 is that  $H_i$  is preferred when  $BF_{ic}$  is larger than 1, and  $H_c$  is preferred when  $BF_{ic}$  is smaller than 1 (Weiss, 1997; Klugkist et al., 2014). In Figure 1a, the vertical line at  $BF_{ic} = 1$  indicates the decision criterion used in this approach: obtaining  $BF_{ic} > 1$  results in the decision that the data support  $H_i$ , and  $BF_{ic} < 1$  results in the decision that the data support  $H_c$ .

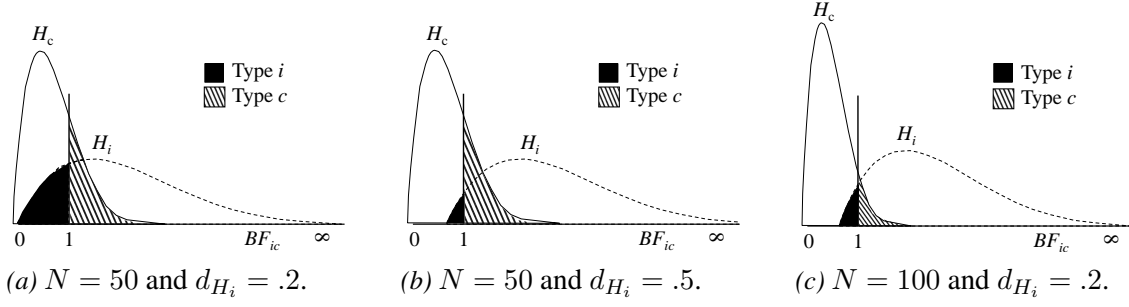


Figure 1. Error probabilities for Approach 1. Hypothetical sampling distributions of  $BF_{ic}$  under  $H_i$  and  $H_c$ , given sample size  $N$  and effect sizes  $d_{H_i}$  and  $d_{H_c}$ . Note that  $d_{H_c} = .2$  in each figure.

The vertical line marks two error probabilities. The first, the probability of observing  $BF_{ic} < 1$  when  $H_i$  is true,  $P(BF_{ic} < 1|H_i)$ , is the probability of supporting  $H_c$  when  $H_i$  is true. In the remainder of this paper, this probability will be referred to as a Type  $i$  error probability. The second error probability is that of observing  $BF_{ic} > 1$  when  $H_c$  is true denoted by  $P(BF_{ic} > 1|H_c)$ , that is, support for  $H_i$  when  $H_c$  is true. This will be referred to as Type  $c$  error probability. The average of Type  $i$  and Type  $c$  error probabilities will be called the *Decision error probability* which is similar to the average error probability used in Reyes and Ghosh (2013).

As can be seen in Figure 1b, if the effect size under  $H_i$  in Figure 1a increases, the sampling distribution under  $H_i$  shifts further away from the decision criterion, thus the Type  $i$  error decreases. As can be seen in Figure 1c, if the group sample size in Figure 1a increases, both Type  $i$  and Type  $c$  error decrease in this situation. For Approach 1, sample size will be determined such that the Type  $i$ , Type  $c$ , or Decision error probability is acceptably low.

### 3.2 Approach 2

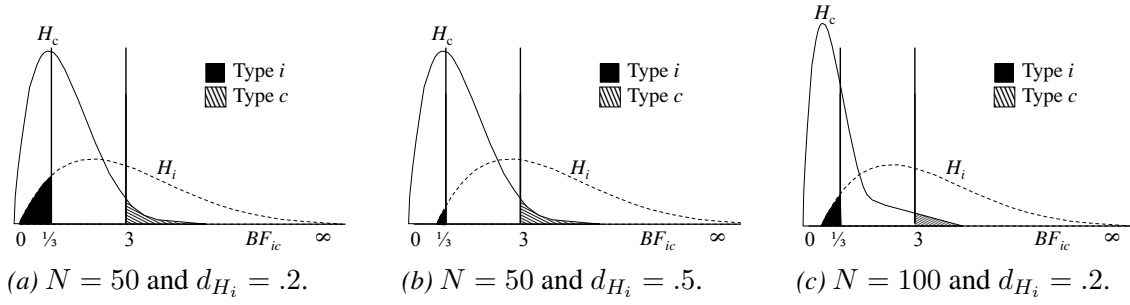
The decision criterion used in Approach 2 allows for indecision. Kass and Raftery (1995) have argued that Bayes factors between  $\frac{1}{3}$  and 3 express too little support to prefer either hypothesis. In Approach 2, like De Santis (2004, 2007), this distinction is used by deciding that  $H_i$  is preferred for Bayes factors larger than 3 and deciding that  $H_c$  is preferred for Bayes factors smaller than  $\frac{1}{3}$ . For Approach 2, Type  $i$  error probability is expressed by  $P(BF_{ic} < \frac{1}{3}|H_i)$  and Type  $c$  error probability by  $P(BF_{ic} > 3|H_c)$ . The average of Type  $i$  and Type  $c$  is the Decision error probability. An additional probability in this approach is that of not making a decision:

$$P(\frac{1}{3} < BF_{ic} < 3) = \frac{P(\frac{1}{3} < BF_{ic} < 3|H_i) + P(\frac{1}{3} < BF_{ic} < 3|H_c)}{2},$$

which is called the *Indecision probability*.

Figure 2 shows hypothetical sampling distributions of  $BF_{ic}$  under  $H_i$  and  $H_c$  and the





**Figure 2.** Error probabilities for Approach 2. Hypothetical sampling distributions of  $BF_{ic}$  under  $H_i$  and  $H_c$ , for sample size  $N$  and effect sizes  $d_{H_i}$  and  $d_{H_c}$ . Note that  $d_{H_c} = .2$  in each figure. The average of the area between  $BF_{ic} = \frac{1}{3}$  and  $BF_{ic} = 3$  under  $H_i$  and the area between  $\frac{1}{3}$  and  $BF_{ic} = 3$  under  $H_c$ , is the Indecision probability.

error probabilities under Approach 2. As can be seen in Figure 2b, if the effect size under  $H_i$  in Figure 2a is increased, the Type  $i$  error probability decreases, while the Type  $c$  error probability remains constant. In Figure 2b it can also be seen that the Indecision probability decreases with the increased effect size. As can be seen in Figure 2c, if the sample size in Figure 2a is increased, the Type  $i$  and Type  $c$  error probabilities decrease. Since for both distributions, the size of the area between  $\frac{1}{3}$  and 3 decreases, the Indecision probability also decreases. For Approach 2, sample size will be determined such that the Type  $i$ , Type  $c$ , or the Decision error probability is acceptably low. Based on the determined sample size and the decision criterion Indecision probability can be computed, but not controlled.

### 3.2.1 Approach 2b

Note that the Indecision probability can be quite large in Approach 2, which might be undesirable for a researcher. Therefore, the situation in which a researcher wants to determine sample size such that the Indecision probability is acceptably low is also considered. We will refer to this approach by Approach 2b. In contrast to Approach 2, for Approach 2b sample size is determined such that the Indecision probability is controlled. Based on the sample size and decision criterion, the error probabilities can be determined, but not controlled.

### 3.3 Approach 3

Approach 3 is different from Approach 1 and 2, because it does not rely on error probabilities or on a fixed decision criterion. In the sampling distributions under  $H_i$  and under  $H_c$  the median Bayes factor can be determined. These medians are an indication of the size of the Bayes factors that can be expected, given  $N$ ,  $d_{H_i}$ , and  $d_{H_c}$ . The median was used, because it has the nice interpretation that exactly 50% of the distribution of Bayes factors is larger than the median, and 50% is smaller.

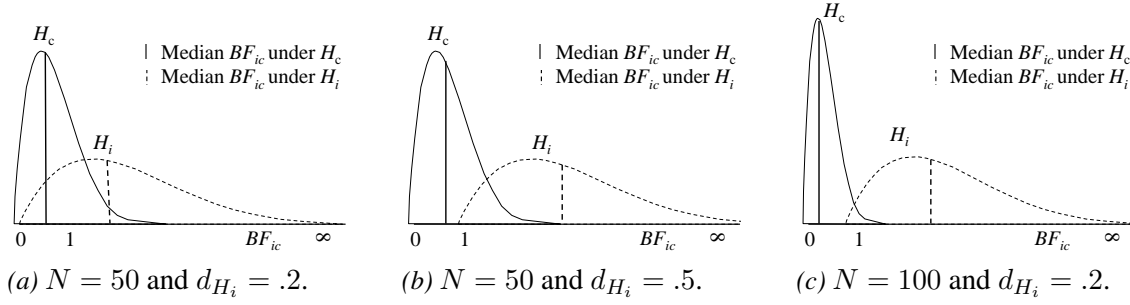


Figure 3. Median Bayes factors for Approach 3. Hypothetical sampling distributions of  $BF_{ic}$  under  $H_i$  and  $H_c$ , given sample size  $N$  and effect size  $d_{H_i}$ . Note that  $d_{H_c} = .2$  in each figure.

Figure 3 shows hypothetical sampling distributions of  $BF_{ic}$  under  $H_i$  and  $H_c$ . As can be seen in Figure 3a, each of the distributions is marked with a line, indicating the median value of that distribution. Note that in Approach 3, a researcher can choose a required value for the median Bayes factor under  $H_i$  or under  $H_c$ . As can be seen in Figure 3b, if the effect size in Figure 3a increases, the median Bayes factor under  $H_i$  increases, while the median Bayes factor under  $H_c$  remains constant. As can be seen in Figure 3c, if the sample size in Figure 3a increases, the median Bayes factor under  $H_i$  increases, while the median Bayes factor under  $H_c$  decreases. For Approach 3, sample size will be determined such that the median Bayes factor under  $H_i$  is of a required size,  $B$ , or the median Bayes factor under  $H_c$  is of a required size,  $1/B$ .

### 3.4 Critical values

Table 1 displays the critical values for the error probabilities, Indecision probability, and median Bayes factor considered in this paper. Note that traditionally in null hypothesis significance testing, Type I and Type II error probabilities are usually set at .05 and .2, resulting in an average error probability of .125. By limiting ourselves to Decision error probabilities of .1, .05, and .025 for Approach 1 and 2 (see Table 1), relatively strict cut-off values are used. We chose to do so, to respond to the replication crisis in social sciences. This crisis is partially due to publication of false positives (see for example Pashler and Wagenmakers (2012) and Thompson (2004)), which are partly caused by too lenient Type I error rates. By using strict error probabilities, we determine sample sizes that have a relatively high probability of rendering correct results. For the Indecision probability in Approach 2b,3, .2, and .1 are considered. Indecision probabilities larger than .3 have not been considered because then studies remain undecided too often. Furthermore, Indecision probabilities smaller than .1 were not considered, because then the Indecision probability becomes too small, and the situation resembles Approach 1 too much.

In Approach 3, 3, 10, and 20 are considered for  $B$ , roughly based on an indication of strength of support by Kass and Raftery (1995). A  $B$  of 3 implies a required median

Table 1  
Critical error probabilities and critical  $B$  considered

Approach		Critical values		
1 and 2	Error probability	.1	.05	.025
2b	Indecision probability	.3	.2	.1
3	$B$	3	10	20

Bayes factor of 3 if  $H_i$  is true, and implies a required median Bayes factor of  $1/B = 1/3$  if  $H_c$  is true. Note that a researcher could decide that both the Bayes factor if  $H_i$  is true and the Bayes factor if  $H_c$  is true, should be of a required size. This is done by determining the sample size such that the median Bayes factor under  $H_i$  is  $B$ , and the sample size such that the median Bayes factor under  $H_c$  is  $1/B$ . The largest of these two sample sizes is the required sample size.

## 4 Methods

Sample size tables are determined through simulations. The simulations are programmed and carried out in R (R Core Team, 2013). The hypotheses considered in this paper are  $H_i$ ,  $H_c$ , and  $H_{i'}$ , like in Equations 3–5, with  $K = 2, 3, 4$ . The R code computes  $BF_{ic}$  or  $BF_{ii'}$ , based on samples from populations under  $H_i$  and  $H_c$  or under  $H_i$  and  $H_{i'}$ . The first three subsections describe in detail how the populations under  $H_i$ ,  $H_c$ , and  $H_{i'}$  are specified. These are the first steps of the simulation procedure. Section 4.4 gives a brief description of the entire simulation procedure by means of an example.

### 4.1 Specify $H_i$ and effect size $d_{H_i}$

First, a population under  $H_i$  needs to be specified. The population is dependent on the number of groups under  $H_i$ , and on effect size  $d_{H_i}$ . As was indicated before, the effect size considered in this paper is Cohen's  $d$ . Based on Cohen's definition of small, medium, and large effect sizes,  $d_{H_i}$  can take on the values 0.2, 0.5, and 0.8 (Cohen, 1992). The group standard deviation  $\sigma_k$  is 1, for  $k = 1, 2, \dots, K$ , and the smallest ordered mean is equal to 0. The difference between the first and the last ordered mean is described by  $d_{H_i}$ , and intermediate means are equally spaced between 0 and  $d_{H_i}$ . Table 2 shows the population means for  $K = 2, 3, 4$ . If  $H_i$  is compared to  $H_c$ ,  $d_{H_i} = .2, .5$ , and  $.8$  are considered. If  $H_i$  is compared to  $H_{i'}$ ,  $d_{H_i} = .2$  and  $.5$  are considered.

Note that because of our definition of effect size, the difference between each pair of means in a hypothesis for some effect size, varies over  $K$ . For example, for  $K = 3$ , and  $d_{H_i} = .2$ , the standardized difference between each pair of means is  $.1$ , while for  $K = 4$ , the difference is  $.067$ . We believe that by controlling the effect size over the difference between the first and the last mean, realistic mean orderings can be expressed. For example, for  $K = 4$ , it would be unrealistic to consider an effect size of  $.8$  between

Table 2  
Population means given  $d$

$K$	$d$	$\mu_1$	$\mu_2$	$\mu_3$	$\mu_4$
2	0.2	0.2	0	-	-
	0.5	0.5	0	-	-
	0.8	0.8	0	-	-
3	0.2	0.2	0.1	0	-
	0.5	0.5	0.25	0	-
	0.8	0.8	0.4	0	-
4	0.2	0.2	0.133	0.067	0
	0.5	0.5	0.333	0.167	0
	0.8	0.8	0.533	0.267	0

Note.  $d$  can be  $d_{H_i}$ ,  $d_{H_c}$ , or  $d_{H_{i'}}$ . The means are labelled such that they match the ordering of means in  $H_i$ . The labels can be rearranged such that they match  $H_c$  or  $H_{i'}$ . For example, if  $K = 3$ ,  $d_{H_{i'}} = .2$ , and  $H_{i'} : \mu_3 > \mu_2 > \mu_1$ , the populations means will be  $\mu_3 = .2$ ,  $\mu_2 = .1$ , and  $\mu_1 = 0$ .

each pair of means, because it would result in a standardized difference of 2.4 between the first and the last ordered mean. Although we believe our choices for effect size are realistic, we also acknowledge that we are being strict by considering rather small differences between pairs of means like .067.

#### 4.2 Specify $H_c$ and effect size $d_{H_c}$

If  $H_i$  is evaluated with  $H_c$ , a population under  $H_c$  needs to be specified. The hypothesis  $H_c$  is the complement of  $H_i$ , indicating that every ordering of means not in  $H_i$  can be true. For  $K = 2$ , only one other ordering than that under  $H_i$  is possible, but five orderings are possible for  $K = 3$ , and 23 for  $K = 4$ . Table 3 shows all options of ordered means under  $H_c$  for  $K = 2, 3$ , and three examples for  $K = 4$ . As can be seen for  $K = 3$ , the orderings violate  $H_i$  in different ways. These violations are classified as small, medium, and large. An example of a small violation is a change in the order of only one pair of means, and an example of a large violation is a complete reversal of the ordering of means under  $H_i$ .

If a researcher is comparing  $H_i$  and  $H_c$ , he is testing an informative hypothesis  $H_i$  against its complement  $H_c$ , that is, he is testing one theory. The required sample size should be such that it can detect any deviation from his theory that is possible under  $H_c$ . Thus, additionally to a small effect size, researchers should always consider small violations under  $H_c$ . For a complete overview, this paper does present sample sizes required for medium and large violations, too.

Only  $d_{H_c} = 0.2$  is considered. By doing so, the required sample sizes are sufficient to detect small deviations from  $H_i$ . We assume that if a researcher wants to evaluate  $H_i$  with  $H_c$ , he wants to be able to detect any deviation from his theory, specified in  $H_i$ . A small effect size under  $H_c$  renders a sample size sufficient to detect small deviations.

Table 3  
Examples of ordered population means under  $H_c$

$K$	Ordering	Violation of $H_i$
2	$\mu_2 > \mu_1$	-
3	$\mu_1 > \mu_3 > \mu_2$	small *
	$\mu_2 > \mu_1 > \mu_3$	small
	$\mu_2 > \mu_3 > \mu_1$	medium *
	$\mu_3 > \mu_1 > \mu_2$	medium
	$\mu_3 > \mu_2 > \mu_1$	large*
4	$\mu_1 > \mu_2 > \mu_4 > \mu_3$	small *
	$\mu_2 > \mu_3 > \mu_1 > \mu_4$	medium *
	$\mu_4 > \mu_3 > \mu_2 > \mu_1$	large *

Note. For  $K = 4$  only a selection of ordered means is presented. A \* indicates that this ordering is used under  $H_{i'}$ . Note that  $d_{H_c} = .2$ , and  $d_{H_{i'}} = .2$  or  $.5$ .

### 4.3 Specify $H_{i'}$ and effect size $d_{H_{i'}}$

If  $H_i$  is evaluated with  $H_{i'}$ , a population under  $H_{i'}$  needs to be specified. To specify a population under  $H_{i'}$ , first a choice needs to be made for what ordering of means is considered under  $H_{i'}$ . Any ordering of means that is possible under  $H_c$  could be used as  $H_{i'}$ . In this paper, one ordering of means with a small violation of  $H_i$  is considered, one with a medium violation, and one with a large violation, for  $K = 3, 4$ . In Table 3 the orderings considered for  $H_{i'}$  are marked with an asterisk.

If  $H_i$  is compared with  $H_{i'}$ ,  $.2$  and  $.5$  are considered for both  $d_{H_i}$  and  $d_{H_{i'}}$ . We do so, because if a researcher wants to evaluate  $H_i$  with  $H_{i'}$ , he might value these two hypotheses equally. He can expect that a population under  $H_i$  is true, with for example an effect size of  $.5$ , but at the same time also consider a population under  $H_{i'}$ , with an effect size of  $.5$ .

### 4.4 Simulation procedure

This section describes the steps taken in the simulation procedure by means of an example. Figure 4 displays the simulation procedure, and highlights the choices made in the example.

1. Specify  $K$ , the number of groups, and the informative hypotheses considered:  $H_i$ , and  $H_c$  or  $H_{i'}$ . For this example,  $K = 3$ ,  $H_i : \mu_1 > \mu_2 > \mu_3$ , which is compared with  $H_c : \text{not } H_i$ .
2. Specify the effect sizes:  $d_{H_i}$  and  $d_{H_c}$  or  $d_{H_{i'}}$ . For this example,  $d_{H_i} = .2$  and  $d_{H_c} = .2$ .
3. Determine the population means based on  $K$ , the effect sizes, and the hypotheses. As indicated before, throughout this paper, the group standard deviation is set to 1.

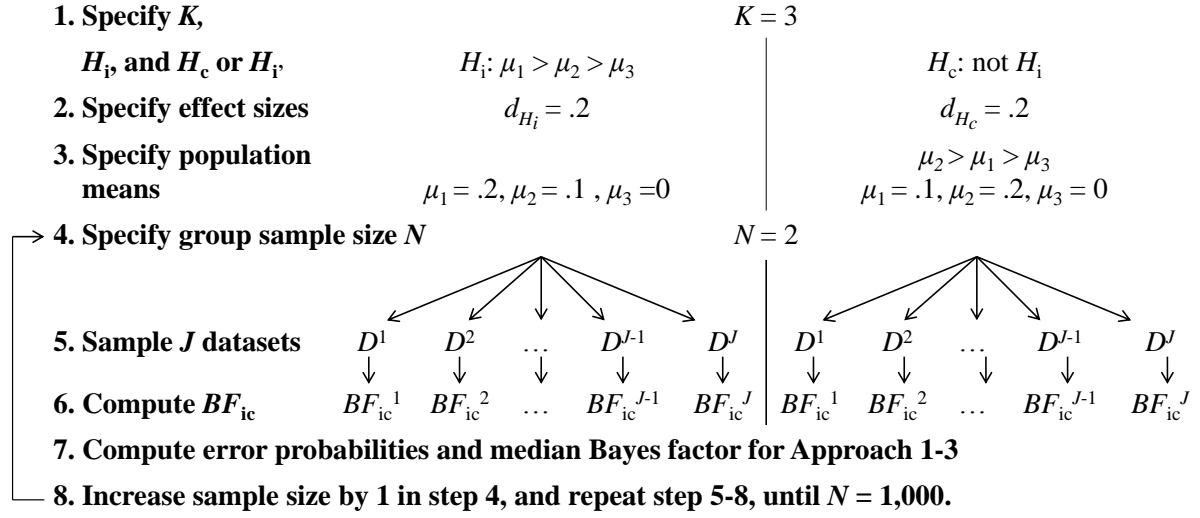


Figure 4. Example of the simulation procedure.

If  $H_c$  is considered, the simulations have to be run in turn for each population mean ordering possible under  $H_c$ , using the same  $K$ , hypotheses, and effect sizes. This continues until all orderings have been considered. In the example, the second ordering from Table 3 is considered. The population means follow from Table 2.

4. Specify a starting group sample size  $N$ . In this example and for all simulations the starting group sample size is 2.
5. Sample  $J$  datasets using the population means and standard deviation, and sample size  $N$ . For all simulations,  $J = 10,000$ .
6. Compute the complexity and fit using Equation 14–15. Compute  $BF_{ic}$  or  $BF_{ii'}$ , using Equation 7 or 8. Since  $H_i$  is compared with  $H_c$  in this example,  $BF_{ic}$  is computed.
7. Compute the appropriate error probabilities, Indecision probability and the median Bayes factor for Approaches 1–3, based on the Bayes factors. Note that the Type  $i$  and Type  $c$  error probabilities are computed separately for Approach 1 and Approach 2, because of the different decision criteria (see Figure 1–3).
8. Increase the group sample size in Step 4 by 1, until  $N = 1,000$ , and repeat Steps 5–8.

Based on the simulations the error probabilities and median Bayes factors for every group sample size from  $N = 2$  to  $N = 1,000$  are known. The required sample size can be determined based on the type and size of error one is willing to make (Approaches 1, 2, and 2b), or on the median Bayes factor (Approach 3). The critical error probabilities and median Bayes factors used are those presented in Table 1. If  $H_c$  is considered, the required sample size is determined for each of the orderings. Then, the orderings are

grouped by violation size, and the average for each of these groups is computed. Thus, if two orderings exist with a small violation size, the average of the required sample sizes for these orderings is the required sample size for small violations.

## 5 Results

Tables 5–18 contain the required group sample sizes based on the simulations for each of the approaches. Separate tables are presented for the comparison of  $H_i$  and  $H_c$ , and for the comparison of  $H_i$  and  $H_{i'}$ . The tables and the examples can contain very small required sample sizes. Note that it is advised to use a group sample size of at least 10, even if a table or an example suggests a smaller sample size. We provide a minimum, because inferences based on small sample sizes are susceptible to outliers. Sections 5.1–5.3 illustrate by means of brief examples how each table can be used. Section 5.4 interprets these tables. Note that because Sections 5.1–5.3 are to some extent repetitive, it may very well be that you want to skip to Section 5.4.

### 5.1 Approach 1

Tables 5 and 6 show the required group sample sizes using Approach 1, with  $K = 2, 3, 4$ , for the evaluation of  $H_i$  and  $H_c$  with  $BF_{ic}$ , and with  $K = 3, 4$ , for the evaluation of  $H_i$  and  $H_{i'}$  with  $BF_{ii'}$ .

*Example 1.1* Suppose a researcher wants to evaluate  $H_i$  with  $H_c$ , with  $K = 3$ . The researcher wants to control the Decision error probability at .05. He specifies  $d_{H_i} = .5$ ,  $d_{H_c} = .2$ , and expects even small violations to be possible under  $H_c$ . As can be seen in Table 5, the required sample size is 977. Suppose this researcher did not consider  $H_c$ , but  $H_{i'}$  with a small deviation of  $H_i$ . As can be seen in Table 6, the required sample size is 327.

*Example 1.2* Suppose a researcher wants to evaluate  $H_i$  with  $H_c$ , with  $K = 3$ . The researcher wants to control the Type  $c$  error probability at .1. He specifies  $d_{H_i} = .2$ ,  $d_{H_c} = .2$ , and expects that only large violations are possible under  $H_c$ . As can be seen in Table 5, the required sample size is 54. Suppose this researcher did not consider  $H_c$ , but  $H_{i'}$  with a large deviation of  $H_i$ . As can be seen in Table 6, the required sample size is 81.

*Example 1.3* Suppose a researcher wants to evaluate  $H_i$  with  $H_c$ , with  $K = 4$ . The researcher wants to control the Type  $i$  error probability at .025. He specifies  $d_{H_i} = .2$ ,  $d_{H_c} = .2$ . As can be seen in Table 5, the required sample size is 443. Suppose this researcher did not consider  $H_c$ , but  $H_{i'}$ . As can be seen in Table 6, the required sample size is larger than 1,000 if  $H_{i'}$  was specified with a small violation of  $H_i$ , the sample size is 575 with a medium violation, and 169 with a large violation.

## 5.2 Approach 2

Tables 7 and 8 show the required group sample sizes using Approach 2, with  $K = 2, 3, 4$ , for the evaluation of  $H_i$  and  $H_c$  with  $BF_{ic}$ , and with  $K = 3, 4$ , for the evaluation of  $H_i$  and  $H_{i'}$  with  $BF_{ii'}$ . Tables 9 and 10 present the corresponding Indecision probabilities.

*Example 2.1* Suppose a researcher wants to evaluate  $H_i$  with  $H_c$ , with  $K = 3$ . The researcher wants to control the Decision error probability at .05. He specifies  $d_{H_i} = .5$ ,  $d_{H_c} = .2$ , and expects even small violations to be possible under  $H_c$ . As can be seen in Tables 7 and 9, the required sample size is 451, and the Indecision probability is .200. Suppose this researcher did not consider  $H_c$ , but  $H_{i'}$  with a small deviation of  $H_i$ . As can be seen in Table 8 and 10, the required sample size is 64, and the Indecision probability is .363.

*Example 2.2* Suppose a researcher wants to evaluate  $H_i$  with  $H_c$ , with  $K = 3$ . The researcher wants to control the Type  $c$  error probability at .1. He specifies  $d_{H_i} = .2$ ,  $d_{H_c} = .2$ , and expects that only large violations are possible under  $H_c$ . As can be seen in Table 7 and 9, the required sample size is 3, and the Indecision probability is .435. Suppose this researcher did not consider  $H_c$ , but  $H_{i'}$  with a large deviation of  $H_i$ . As can be seen in Table 8 and 10, the required sample size is 34, and the Indecision probability is .256.

*Example 2.3* Suppose a researcher wants to evaluate  $H_i$  with  $H_c$ , with  $K = 4$ . The researcher wants to control the Type  $i$  error probability at .025. He specifies  $d_{H_i} = .2$ ,  $d_{H_c} = .2$ . As can be seen in Table 7 and 9, the required sample size is 233, and the Indecision probability is .160. Suppose this researcher did not consider  $H_c$ , but  $H_{i'}$ . As can be seen in Table 8 and 10, the required sample size is 628 if  $H_{i'}$  was specified with a small violation of  $H_i$  with an Indecision probability of .429, and the required sample size is 318 with a medium violation with an Indecision probability of .265, and a sample size of 114 with a large violation with an Indecision probability of .144.

### 5.2.1 Approach 2b

Tables 11 and 12 show the required group sample sizes using Approach 2b, with  $K = 2, 3, 4$ , for the evaluation of  $H_i$  and  $H_c$  with  $BF_{ic}$ , and with  $K = 3, 4$ , for the evaluation of  $H_i$  and  $H_{i'}$  with  $BF_{ii'}$ . Tables 13 and 14 depict the corresponding Type  $i$ , Type  $c$  or Type  $i'$ , and Decision error probability.

*Example 2b.1* Suppose a researcher wants to evaluate  $H_i$  with  $H_c$ , with  $K = 3$ . The researcher wants to control the Indecision probability at .2. He specifies  $d_{H_i} = .5$ ,  $d_{H_c} = .2$ , and expects even small violations to be possible under  $H_c$ . As can be seen in Table 11 and 13, the required sample size is 427, and the Type  $i$  error probability is smaller than .001, Type  $c$  is .112, and the Decision error probability is .056. Suppose this researcher did not consider  $H_c$ , but  $H_{i'}$  with a small deviation of  $H_i$ . As can be seen in Table 12 and 14, the required sample size is 190, and the Type  $i$  error probability is .001, Type  $c$  is .044, and the Decision error probability is .023.

*Example 2b.2* Suppose a researcher wants to evaluate  $H_i$  with  $H_c$ , with  $K = 3$ .



The researcher wants to control the Indecision probability at .1. He specifies  $d_{H_i} = .2$ ,  $d_{H_c} = .2$ , and expects that only large violations are possible under  $H_c$ . As can be seen in Table 11 and 13, the required sample size is 281, and the Type  $i$  error probability is .008, Type  $c$  is smaller than .001, and the Decision error probability is .004. Suppose this researcher did not consider  $H_c$ , but  $H_{i'}$  with a large deviation of  $H_i$ . As can be seen in Table 12 and 14, the required sample size is 61, and the Type  $i$  error probability is .059, Type  $c$  is smaller than .001, and the Decision error probability is .030.

*Example 2b.3* Suppose a researcher wants to evaluate  $H_i$  with  $H_c$ , with  $K = 4$ . The researcher wants to control the Indecision probability at .3. He specifies  $d_{H_i} = .2$ ,  $d_{H_c} = .2$ , and expects medium violations of  $H_i$  under  $H_c$ . As can be seen in Table 11 and 13, the required sample size is 55, and the Type  $i$  error probability is .012, Type  $c$  is .118, and the Decision error probability is .065. Suppose this researcher did not consider  $H_c$ , but  $H_{i'}$  with a medium violation of  $H_i$ . As can be seen in Table 12 and 14, the required sample size is 2 and the Type  $i$  error probability is .231, Type  $c$  is .383, and the Decision error probability is .307.

### 5.3 Approach 3

Table 15–18 show the required group sample sizes for  $K = 2, 3, 4$ , for the evaluation of  $H_i$  and  $H_c$  with  $BF_{ic}$ , using Approach 3. Tables 15 and 16 show the required group sample sizes if the median  $BF_{ic}$  or  $BF_{ii'}$  is required to be of size  $B$  under  $H_i$ , and Tables 17 and 18 show the required group sample sizes if the median  $BF_{ic}$  or  $BF_{ii'}$  is required to be of size  $1/B$  under  $H_c$  or  $H_{i'}$ .

*Example 3.1* Suppose a researcher wants to evaluate  $H_i$  with  $H_c$ , with  $K = 3$ . The researcher wants to know that 50% of the possible Bayes factors if  $H_i$  is true, is larger than 10, and that 50% of the possible Bayes factors if  $H_c$  is true, is smaller than  $\frac{1}{10}$ . He specifies  $d_{H_i} = .5$ ,  $d_{H_c} = .2$ , and expects even small violations to be possible under  $H_c$ . As can be seen in Tables 15 and 17, the required sample sizes to meet the boundaries are 50 and 839, respectively. Because the researchers wants to adhere to both boundaries, the largest sample size is required, which is 839. Additionally, we know that 13.4% of the possible Bayes factors under  $H_c$  is larger than 1, which implies that the probability of finding evidence in favour of the  $H_i$  when  $H_c$  is true, is .134.

Suppose this researcher did not consider  $H_c$ , but  $H_{i'}$  with a small deviation of  $H_i$ . As can be seen in Table 16 and 18, the required samples size are 63 and 391, respectively. Again, because both boundaries need to be adhered, the largest sample size must be considered, which is 391. Additionally, we know that 7.8% of the possible Bayes factors under  $H_c$  is larger than 1, thus a probability of .078 to find evidence in favour of  $H_i$  when  $H_c$  is true.

*Example 3.2* Suppose a researcher wants to evaluate  $H_i$  with  $H_c$ , with  $K = 3$ . The researcher wants to know that 50% of the possible Bayes factors if  $H_c$  is true, is smaller than  $\frac{1}{3}$ . He specifies  $d_{H_i} = .2$ ,  $d_{H_c} = .2$ , and expects that only large violations are possible under  $H_c$ . As can be seen in Table 17, the required sample size is 2, and additionally,

we know that 30.3% of possible Bayes factors under  $H_c$  is larger than 1. This researcher has a probability of .303 to find evidence in favour of  $H_i$  when  $H_c$  is true. Suppose this researcher did not consider  $H_c$ , but  $H_{i'}$  with a large deviation of  $H_i$ . As can be seen in Table 18, the required sample size is 9, and additionally, we know that 33.2% of possible Bayes factors under  $H_{i'}$  is larger than 1. This researcher has a probability of .332 to find evidence in favour of  $H_i$  when  $H_c$  is true.

*Example 3.3* Suppose a researcher wants to evaluate  $H_i$  with  $H_c$ , with  $K = 4$ . The researcher wants to know that 50% of possible Bayes factors if  $H_i$  is true, is larger than 20. He specifies  $d_{H_i} = .2$ ,  $d_{H_c} = .2$ . As can be seen in Table 15 the required sample size is 637, and additionally, we know that 1.1% of all possible Bayes factors under  $H_i$  is smaller than 1. Suppose this researcher did not consider  $H_c$ , but  $H_{i'}$ . As can be seen in Table 16 the required sample size is 38 using a large violation of  $H_i$  under  $H_{i'}$ , with 18% of possible Bayes factors under  $H_i$  smaller than 1. For medium violations, the sample size is 283, with 9.1% of possible Bayes factors smaller than 1, and for small violations, the sample size required is larger than 1,000.

#### 5.4 Discussion of table features

This section discusses two features of the tables presented in Section 5. First, the required sample sizes are compared to sample sizes presented by Cohen (1992) for the evaluation of  $H_0$  and  $H_1$ . Secondly, the benefit of using  $H_{i'}$  over  $H_c$  is discussed.

The sample sizes presented in Tables 5–18 might seem large on first view. However, in this paper strict measures for the effect sizes and the error probabilities have been used. Small, medium, and large effect sizes are used, however, these effect sizes describe the difference between the largest and the smallest mean. Thus, large differences between each pair of means are not common. As was explained in Section 3.4, the used critical values in this paper (.1, .05, and .025) are more strict than the Decision error probability based on the traditional Type I and Type II error probabilities  $((.05 + .2)/2 = .25/2 = .125)$ .

In order to put the results obtained in this paper in perspective, the sample sizes based on the approaches in this paper are compared with the sample sizes presented by Cohen (1992). Table 4 presents required sample sizes for  $K = 2$ , using Cohen's comparison of  $H_0$  and  $H_1$ , and for each of the approaches presented in this paper. Specifically, all approaches are compared to the sample sizes for the evaluation of  $H_0 : \mu_1 = \mu_2$  and  $H_1 : \mu_1 \neq \mu_2$ , with a Type I error probability of .05, and a Type II error probability of .80, that is, a Decision error probability of .125. Only  $K = 2$  is considered, because for  $K = 2$ , the effect sizes used in this paper corresponds to the effect sizes used by Cohen. Furthermore, although the comparison of  $H_i : \mu_1 > \mu_2$  with  $H_c : \mu_1 < \mu_2$  is different from the comparison of  $H_0$  with  $H_1$ , it does give an impression of how the required sample sizes compare. For each approach, the type and size of the critical value is specified such that the method is as similar as possible to the Decision error probability used by Cohen. A Decision error probability under Approach 1 is most comparable to that

Table 4  
Required sample sizes

Approach	Critical value		Effect size		
	Type	Size	.2	.5	.8
Cohen (1992)	Decision error	.125	393	64	26
Approach 1	Decision error	.100	82	40	36
Approach 2	Decision error	.100	20	8	6
	<i>Indecision</i>		.422	.384	.343
Approach 2b	Indecision	.100	182	108	108
	<i>Decision error</i>		.005	.007	.007
Approach 3	$B$ under $H_i$	10	87	15	6
	$P(BF_{ic} < 1 H_i)$		.091	.089	.091
Approach 3	$1/B$ under $H_c$	$\frac{1}{10}$	87		
	$P(BF_{ic} > 1 H_c)$		.091		

*Note.* Effect size indicates  $d$  for Cohen (1992), and  $d_{H_i}$  for Approach 1–3. Note that  $d_{H_c} = .2$  for all approaches. Note that Cohen’s approach compares  $H_0$  and  $H_1$ , while Approaches 1–3 compare  $H_i$  and  $H_c$ . Entries in italics are additional probabilities rendered by an approach.

of Cohen when it is .100, since larger values than .100 are not considered in this paper. A Decision error probability under Approach 2 is most comparable to that of Cohen when it is .100. Approach 2b seemed most comparable to Cohen when an Indecision probability of .100 is considered, since additionally to the Indecision probability, this approach renders Decision error probability. For Decision error probabilities smaller than .025, Approach 2b results in a higher probability of a correct decision than Cohen’s setup. Finally, in Approach 3  $B$  does not correspond to a Decision error of a certain size. Therefore,  $B = 10$  is considered, which is usually considered to express a fair amount of evidence.

As can be seen in Table 4, for a small effect size under  $H_i$  and  $H_c$ , the required sample size comparing  $H_i$  and  $H_c$  with Approaches 1–3 is smaller than required for comparing  $H_0$  and  $H_1$  using Cohen’s method. Furthermore, for all approaches but Approach 2b, the required sample size for a medium effect size is smaller than that required for Cohen’s approach.

The comparison with the sample sizes prescribed for null hypothesis significance testing puts the results of this paper in perspective. For small to medium effect sizes, which are often expected in applied research, Approaches 1–3 require smaller sample sizes than Cohen’s power analysis. For Approach 2, it appears that the smaller sample sizes do come at the cost of a relatively large Indecision probability. For large effect sizes, the required sample sizes are smaller for Approach 2 and 3 relative to Cohen’s

sample size, and for the other approaches, the sample sizes are not much larger than those following from Cohen (1992). For  $K = 3, 4$ , the sample sizes are less easy to compare, because of different uses of effect size, and more complex hypotheses. However, if the results are compared, comparing  $H_i$  to  $H_c$  will require similar sample sizes to Cohen, whereas comparing  $H_i$  to  $H_{i'}$  will in general result in smaller sample sizes.

As can be seen in the tables, in general, a smaller sample size is required if  $H_i$  is compared to  $H_{i'}$  than when it is compared to  $H_c$ . For example, as can be seen in Table 5, the required sample size for  $K = 3$ ,  $d_{H_i} = .5$ , and a Decision error probability of .05, the required sample size is 977 for small violations of  $H_c$ , which is the only violation that should be considered in practice. As can be seen in Table 6, if  $H_{i'}$  is considered, the sample size ranges from 22 to 327, dependent on the choice of violation size and effect size under  $H_{i'}$ . All of these sample sizes are much smaller than the 977 required for the comparison of  $H_i$  to  $H_c$ . Thus, if you have a competing theory, you are better off using  $H_{i'}$  than  $H_c$ . Note that the required sample size is not in all situations smaller when using  $H_{i'}$  rather than  $H_c$ . Appendix A explains situations in which this is not the case, and further elaborates on some numerical characteristics of the tables.

## 6 In practice

This section provides guidelines for applied researchers to select an approach,  $H_{i'}$  or  $H_c$ , an effect size, and a critical value. Figure 5 shows a decision tree, with some example research questions. First of all, the decision tree will be discussed, and then the further choices that must be made.

As can be seen in Figure 5, the choice for an approach depends on maximally two sequential questions. The first question, *What type of decision do you want to make?* relates to whether a dichotomous, trichotomous, or no decision should be made. For dichotomous decisions, that is, choosing between  $H_i$  and  $H_c$  or  $H_{i'}$ , Approach 1 applies. For trichotomous decisions, that is, choosing between  $H_i$ ,  $H_c$  or  $H_{i'}$ , and indecision, either Approach 2 or 2b applies. For situations in which a researcher does not want to make decision, but express the support in the data for each hypothesis, Approach 3 applies. If a trichotomous decision is required, the second question, *What probability do you want to control for?* has to be answered. This relates to whether a researcher wants to control the Indecision probability, that is, Approach 2b, or control the Type  $i$ ,  $c$ ,  $i'$ , or Decision error probability, that is, Approach 2.

*Example 1.* Suppose a researcher wants to see if a new drug is more effective than a placebo,  $H_i : \mu_{\text{new}} > \mu_{\text{placebo}}$ , and compares this with the complement,  $H_c$ . It is very important to know if  $H_i$  or  $H_c$  is true, to support the decision to implement the drug or not. Answering Question 1 in Figure 5 this researcher would need to use Approach 1 to determine the required group sample size, because a dichotomous decision has to be made. (cf. Tables 5–6).

*Example 2.* Suppose a researcher wants to investigate whether flyers or posters are more effective in informing inhabitants of a neighbourhood about upcoming events,  $H_i :$

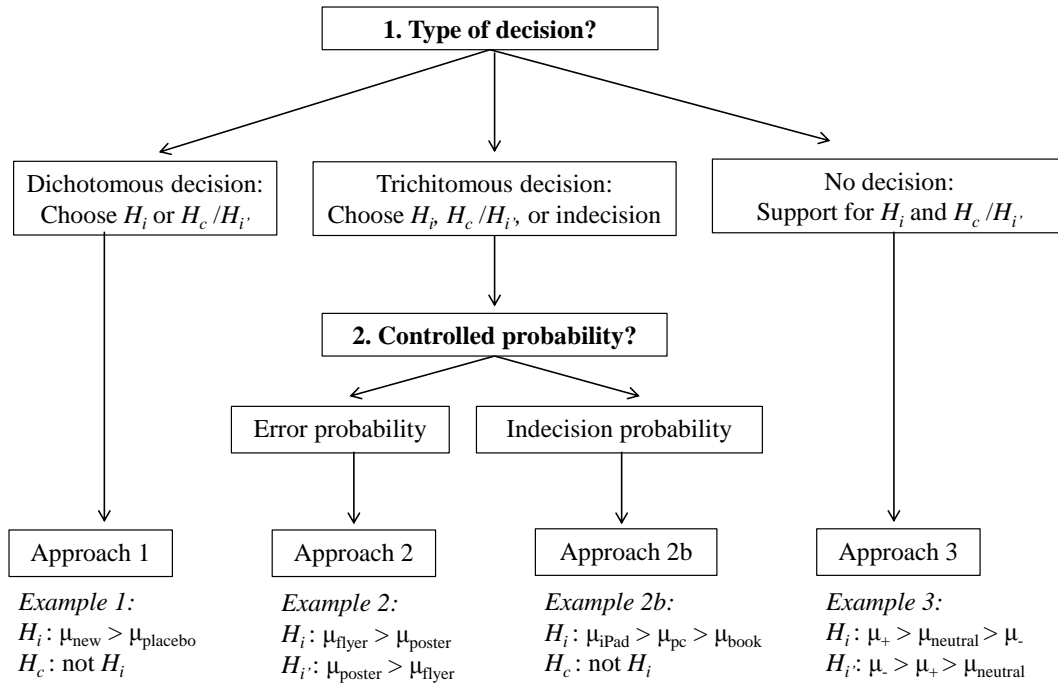


Figure 5. Decision Tree

$\mu_{\text{flyer}} > \mu_{\text{poster}}$  versus  $H_{i'} : \mu_{\text{poster}} > \mu_{\text{flyer}}$ . The researcher wants to make a decision for  $H_i$  or  $H_{i'}$  only when the evidence is sufficiently large. He is open to the fact that the Bayes factor may be too small, and thus replies to Question 1 that he wants to make a trichotomous decision, where he allows for indecision. Finally, he does not have a limit to what indecision he maximally allows, so he replies to Question 2 that he wants to control the error probability. This researcher would need to use Approach 2 to determine the required group sample size (cf. Tables 7–10).

*Example 2b.* Suppose a researcher wants to investigate the effect of learning tool on the test outcome of students. He hypothesizes  $H_i : \mu_{\text{iPad}} > \mu_{\text{PC}} > \mu_{\text{book}}$ , and  $H_c : \text{not } H_i$ . The researcher wants to make a decision for  $H_i$  or  $H_c$  only when the evidence is sufficiently large. He is open to the fact that the Bayes factor may be too small, and thus replies to Question 1 that he wants to make a trichotomous decision, where he allows for indecision. Because his research is quite costly to execute, he wants to limit the Indecision probability. Therefore, this researcher should use Approach 2b to determine the required group sample size (cf. Tables 11–14).

*Example 3.* Suppose a researcher wants to evaluate two competing theories. The theories concern the attitude of people towards healthy food, after being primed with positive, neutral, or negative cues. He hypothesizes  $H_i : \mu_+ > \mu_{\text{neutral}} > \mu_-$  and  $H_{i'} : \mu_- > \mu_+ > \mu_{\text{neutral}}$ . This researcher is not interested in making a decision, but wants to express the support in the data for  $H_i$  and  $H_{i'}$ . Following Question 1 in Figure 5, he needs to use Approach 3 to determine the required sample size (cf. Tables 17–18).

After determining the appropriate approach, a researcher still needs to make three decisions. First of all, a researcher needs to decide whether he wants to compare  $H_i$  to

$H_c$  or  $H_{i'}$ . If  $H_c$  is used, as explained in Section 4.2, only small violations of  $H_i$  should be considered, and if  $H_{i'}$  is used, the researcher must decide based on his theory, what the ordering of means under  $H_{i'}$  is. Table 3 displays what is considered a small violation under  $H_c$ , and shows the orderings considered under  $H_{i'}$  in this paper

Secondly, a researcher needs to choose the effect sizes and population means under  $H_i$  and  $H_c$  or  $H_{i'}$ . Table 2 displays the population means for the effect sizes considered in this paper. Inspiration for effect size can be taken from previous research in the same field. If the effect size generally is .5, use .5. If no previous research exists, it is up to the researcher to choose a reasonable effect size. It is advised to use a small effect size in this situation.

Thirdly, a researcher needs to make one or two decisions regarding the critical value. This differs per approach. Table 1 displays the critical values for the different decision criteria used in this paper. For Approach 1 and 2, a researcher must first decide whether he wants to control Type  $i$ , Type  $c$  or Type  $i'$ , or Decision error probability. This choice is dependent on what type of error the researcher values more strongly. For example, if a Type  $i$  error is deemed most harmful, the Type  $i$  error probability must be controlled. Secondly, the researcher must choose the critical value. This should be done based on practical value. The smaller the value, the larger the probability that the resulting decision will be correct.

For Approach 2b, a researcher must only decide what critical value he considers for the Indecision probability. This choice depends on the costs related to not making a decision. If the costs are high, a small critical value should be chosen for the Indecision probability.

For Approach 3, a researcher must first decide whether he wants to control the median Bayes factor under  $H_i$ , the median Bayes factor under  $H_c$  or  $H_{i'}$ , or control both. For example, if the evidence under  $H_i$  is deemed most important, the chosen  $B$  only refers to Bayes factors under  $H_i$ . Secondly, the researcher must choose a size of this median Bayes factor, which is expressed by  $B$ . This should be done based on practical value. Tentative guidelines for the strength of the evidence expressed by  $B$  can be found in (Kass & Raftery, 1995). According to them,  $B = 3$  expresses positive support, and  $B = 20$  expresses strong support.

## 7 Discussion

In this paper, sample sizes have been determined for the comparison of  $H_i$  with  $H_c$  or  $H_{i'}$ , by means of three main approaches. As was indicated in Section 3 and 4, strict effect sizes and critical values have been used. The effect sizes have been chosen such that they gave a reasonable representation of what can be expected in social sciences. For the error probabilities, it should be noted again that strict error probabilities are required for more sound research outcomes. In order to accommodate researchers that want different effect sizes, error probabilities, or different orderings of means under  $H_{i'}$  than those considered

in this paper, an R script has been developed<sup>†</sup>.

Future research could investigate the required sample sizes for different types of hypotheses. For example, the use of a composite  $H_{i'}$ , that consists of multiple orderings, can be of interest for researchers that consider not all orderings under  $H_c$  relevant, but do not have one specific theory. Furthermore, other informative hypotheses than simple order constrained hypotheses could be considered. Finally, only have been considered ANOVA models. It would be interesting to extend this research to other statistical models.

---

<sup>†</sup>The script and the manual can be downloaded using this link:  
[https://www.dropbox.com/sh/0bfa2pqplfhwfkj/AABizVs\\_0TLWiLcgYvogEwEEa?dl=0](https://www.dropbox.com/sh/0bfa2pqplfhwfkj/AABizVs_0TLWiLcgYvogEwEEa?dl=0)

## References

- Adcock, C. J. (1997). Sample size determination: A review. *Journal of the Royal Statistical Society, Series D*, 46, 261-283.
- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Proc. 2nd Int. Symp. Information Theory* (p. 267-281). Budapest: Akademiai kiado.
- Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. New Jersey: Lawrence Erlbaum.
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112, 155-159.
- Cohen, J. (1994). The earth is round ( $p < .05$ ). *American Psychologist*, 49, 997-1003.
- De Santis, F. (2004). Statistical evidence and sample size determination for Bayesian hypotheses testing. *Journal of Statistical Planning and Inference*, 124, 121-144.
- De Santis, F. (2007). Alternative Bayes factors: Sample size determination and discriminatory power assessment. *Test*, 16, 504-522.
- Gu, X., Mulder, J., Deković, M., & Hoijtink, H. (2014). Bayesian evaluation of inequality constrained hypotheses. *Psychological Methods*, 19, 511-527.
- Hoijtink, H. (2012). *Informative Hypotheses. Theory and Practice for Behavioral and Social Scientists*. Boca Raton: Chapman & Hall/CRC.
- Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773-795.
- Klugkist, I., Post, L., Haarhuis, F., & van Wesel, F. (2014). Confirmatory methods, or huge samples, are required to obtain power for the evaluation of theories. *Open Journal for Statistics*, 4, 710-725.
- Kuiper, R., & Hoijtink, H. (2010). Comparisons of means using exploratory and confirmatory approaches. *Psychological Methods*, 15, 69-86.
- Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, 7, 528-530.
- R Core Team. (2013). R: A language and environment for statistical computing. [Computer software manual]. Vienna, Austria. Retrieved from <http://www.R-project.org/>
- Reyes, E. M., & Ghosh, S. K. (2013). Bayesian average error-based approach to sample size calculations for hypotheses testing. *Journal of Biopharmaceutical Statistics*, 23, 569-588.
- Rozeboom, W. W. (1997). Good science is abductive not hypothetico-deductive. In L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (p. 335-392). Mahwah, NJ: Erlbaum.
- Silvapulle, M. J., & Sen, P. K. (2004). *Constrained statistical inference: Order, inequality and shape constraints*. London: Wiley.
- Thompson, B. (2004). The "significance" crisis in psychology and education. *The Journal of Socio-Economics*, 33, 607-613.



- Vanbrabant, L., van de Schoot, R., & Rosseel, Y. (2015). Constrained statistical inference: sample-size tables for anova and regression. *Frontiers in Psychology*, 5. doi: 10.3389/fpsyg.2014.01565
- Van de Schoot, R., Hoijtink, H., & Romeijn, J. W. (2011). Moving beyond traditional null hypothesis testing: Evaluating expectations directly. *Frontiers in Psychology*, 2. doi: 10.3389/fpsyg.2011.00024
- Weiss, R. (1997). Bayesian sample size calculations for hypothesis testing. *The Statistician*, 46, 185-191.

Table 5  
Required group sample sizes for Approach 1 using  $H_c$

		Error probability											
		.1			.05			.025			.1	.05	.025
$K$	$d_{H_i} =$	.2	.5	.8	.2	.5	.8	.2	.5	.8	-	-	-
2		82	40	36	132	83	82	187	132	132	82	132	187
		83	14	6	132	23	9	187	32	12			
3	$s$	644	624	624	994	<b>977</b>	977	*	*	*	977	*	*
	$m$	136	56	43	222	110	103	313	180	180	103	180	255
	$l$	109	35	24	181	65	58	278	108	100	<b>54</b>	100	148
		159	28	12	258	42	17	361	58	24			
4	$s$	*	*	*	*	*	*	*	*	*	*	*	*
	$m$	353	274	271	547	486	486	749	690	690	486	690	*
	$l$	146	47	31	241	92	78	345	151	142	76	141	206
		207	35	15	314	51	21	<b>443</b>	75	30			



*Note.* Non shaded areas give the sample size needed to control the Decision error probability,  gives the sample size needed to control the Type  $i$  error probability, and  gives the sample size needed to control the Type  $c$  error probability. Let \* denote group sample sizes larger than 1,000. Small, medium, and large violations under  $H_c$  are denoted by  $s$ ,  $m$ , and  $l$ . Note that  $s$  gives the average over the required group sample sizes for all population mean orderings under  $H_c$  that are a small violation of  $H_i$ . This is analogous for medium and large violations. Note that  $d_{H_c} = .2$  for all sample sizes.

Table 6  
Required group sample sizes for Approach 1 using  $H_{i'}$

$K$	$d_{H_{i'}}$	$d_{H_i} =$	Error probability								
			.1		.05		.025		.1	.05	.025
			.2	.5	.2	.5	.2	.5	-	-	-
3	$s$	.2	318	147	531	<b>327</b>	731	531	318	531	731
		.5	147	51	327	88	531	117	51	88	117
		-	318	51	531	88	731	117			
	$m$	.2	103	54	180	108	252	180	108	180	249
		.5	54	18	103	29	180	40	17	30	41
		-	103	18	180	29	252	40			
	$l$	.2	81	41	135	81	192	135	<b>81</b>	135	192
		.5	41	14	81	22	135	31	14	22	31
		-	81	14	135	22	192	31			
4	$s$	.2	725	338	*	725	*	*	725	*	*
		.5	338	115	725	189	*	273	115	189	273
		-	725	115	*	189	*	273			
	$m$	.2	247	111	399	233	562	382	228	382	540
		.5	124	40	257	63	415	91	38	64	88
		-	257	41	415	63	<b>575</b>	91			
	$l$	.2	73	37	124	73	169	124	73	124	169
		.5	36	13	73	20	123	30	13	20	30
		-	73	13	123	20	<b>169</b>	28			

Note. Non shaded areas give the sample size needed to control the Decision error probability,      gives the sample size needed to control the Type  $i$  error probability, and      gives the sample size needed to control the Type  $i'$  error probability. Let \* denote group sample sizes larger than 1,000. Small, medium, and large violations under  $H_{i'}$  are denoted by  $s$ ,  $m$ , and  $l$ .

Table 7  
Required group sample sizes for Approach 2 using  $H_c$

groups	$d_{H_i} =$	Error probability											
		.1			.05			.025			.1	.05	.025
		.2	.5	.8	.2	.5	.8	.2	.5	.8	-		
2		20	8	6	48	23	20	77	50	48	20	48	77
		20	4	2	48	8	4	77	15	6			
3	$s$	175	39	10	459	<b>451</b>	451	730	730	730	451	730	985
	$m$	30	9	5	72	22	14	130	48	36	6	34	78
	$l$	22	7	4	58	17	10	103	30	19	<b>3</b>	14	32
		51	9	5	99	18	8	169	27	12			
4	$s$	*	*	*	*	*	*	*	*	*	*	*	*
	$m$	94	19	8	233	154	146	388	332	329	140	329	494
	$l$	44	11	6	96	24	13	163	46	30	2	23	61
		80	16	7	146	24	12	<b>233</b>	38	17			



*Note.* Non shaded areas give the sample size needed to control the Decision error probability,  gives the sample size needed to control the Type  $i$  error probability, and  gives the sample size needed to control the Type  $c$  error probability. Let \* denote group sample sizes larger than 1,000. Small, medium, and large violations under  $H_c$  are denoted by  $s$ ,  $m$ , and  $l$ . Sample size is determined for each ordering of population means under  $H_c$ , and then averaged and presented in violation categories. Note that  $d_{H_c} = .2$  for all sample sizes.

Table 8  
Required group sample sizes for Approach 2 using  $H_{i'}$

$K$	$d_{H_{i'}}$	$d_{H_i} =$	Error probability								
			.1		.05		.025		.1	.05	.025
			.2	.5	.2	.5	.2	.5	-	-	-
3	$s$	.2	46	17	147	<b>64</b>	291	147	46	147	291
		.5	17	9	64	29	147	49	9	29	49
		-	46	9	147	29	291	49			
	$m$	.2	42	19	86	46	147	88	42	86	143
		.5	18	8	45	16	86	24	8	15	25
		-	42	8	86	17	147	24			
	$l$	.2	34	16	72	35	115	72	<b>34</b>	72	115
		.5	16	7	35	13	72	19	7	13	19
		-	34	7	72	13	115	19			
4	$s$	.2	69	26	313	133	628	327	69	313	628
		.5	26	15	133	51	327	104	15	51	104
		-	69	15	313	51	<b>628</b>	104			
	$m$	.2	76	30	176	87	314	174	68	167	294
		.5	31	14	91	30	189	51	14	30	49
		-	80	14	189	33	<b>318</b>	52			
	$l$	.2	40	19	76	43	114	76	40	76	114
		.5	19	8	43	13	76	20	8	13	20
		-	40	8	76	14	<b>114</b>	20			



Note. Non shaded areas give the sample size needed to control the Decision error probability,  gives the sample size needed to control the Type  $i$  error probability, and  gives the sample size needed to control the Type  $c$  error probability. Let \* denote group sample sizes larger than 1,000. Small, medium, and large violations under  $H_{i'}$  are denoted by  $s$ ,  $m$ , and  $l$ .

Table 9  
Indecision probabilities for Approach 2 using  $H_c$

Error probability														
$K$	$d_{H_i} =$	.1			.05			.025			.1		.05	.025
		.2	.5	.8	.2	.5	.8	.2	.5	.8	.226	.164		
2		.422	.384	.343	.328	.282	.226	.262	.181	.164	.226	.164	.131	
		.422	.423	.377	.328	.384	.367	.262	.332	.343				
3	$s$	.389	.375	.416	.236	.200	.200	.162	.149	.149	.200	.149	.109	
	$m$	.477	.472	.437	.388	.386	.344	.286	.268	.225	.435	.226	.159	
	$l$	.455	.459	.441	.371	.385	.363	.271	.307	.271	.435	.316	.200	
4		.380	.449	.421	.292	.388	.395	.187	.330	.338				
	$s$	*	*	*	*	*	*	*	*	*	*	*	*	*
	$m$	.427	.421	.400	.297	.242	.226	.199	.158	.152	.259	.152	.103	
	$l$	.395	.407	.389	.311	.340	.317	.219	.241	.219	.340	.269	.155	
		.340	.379	.381	.240	.337	.325	.160	.273	.280				

Note. Non shaded areas give the Indecision probability belonging to the sample size needed to control the Decision error probability, gives the Indecision probability belonging to the sample size needed to control the Type  $i$  error probability, and gives the Indecision probability belonging to the sample size needed to control the Type  $c$  error probability. Let \* denote group sample sizes larger than 1,000. Small, medium, and large violations under  $H_c$  are denoted by  $s$ ,  $m$ , and  $l$ . Indecision probability is determined based on the sample size for each ordering of population means under  $H_c$ , and then averaged and presented in violation categories. Note that  $d_{H_c} = .2$  for all sample sizes.

Table 10  
Indecision probability for Approach 2 using  $H_{i'}$

$K$	$d_{H_{i'}}$	$d_{H_i} =$	Error probability							
			.1		.05		.025		.1	
			.2	.5	.2	.5	.2	.5	.1	.05
3	$s$	.2	.519	.506	.401	<b>.363</b>	.289	.233	.531	.467
		.5	.506	.496	.363	.381	.233	.281	.496	.381
		-	.531	.496	.467	.381	.403	.281		
		3	.279	.250	.204	.159	.128	.104	.308	.251
3	$m$	.2	.252	.264	.161	.184	.106	.129	.264	.169
		.5	.304	.264	.265	.189	.221	.123		
		-	.242	.211	.174	.139	.115	.089	<b>.256</b>	.222
		3	.211	.214	.139	.149	.089	.109	.214	.149
4	$l$	.2	.256	.214	.222	.149	.199	.109		
		.5	.593	.566	.428	.390	.298	.245	.592	.512
		-	.566	.557	.389	.420	.245	.293	.557	.420
		4	.592	.557	.512	.420	<b>.429</b>	.293		
4	$m$	.2	.352	.323	.238	.200	.147	.123	.378	.317
		.5	.318	.324	.194	.233	.121	.150	.324	.218
		-	.375	.324	.316	.233	<b>.265</b>	.154		
		4	.183	.147	.120	.090	.080	.061	.187	.168
4	$l$	.2	.147	.152	.090	.111	.061	.071	.152	.111
		.5	.188	.152	.168	.105	<b>.144</b>	.071		
		-								
		4								

*Note.* Non shaded areas give the Indecision probability belonging to the sample size needed to control the Decision error probability, **■** gives the Indecision probability belonging to the sample size needed to control the Type  $i$  error probability, and **■** gives the Indecision probability belonging to the sample size needed to control the Type  $c$  error probability. Let \* denote group sample sizes larger than 1, 000. Small, medium, and large violations under  $H_{i'}$  are denoted by  $s$ ,  $m$ , and  $l$ .

Table 11  
*Required group sample sizes for Approach 2b using  $H_c$*

$K$	$d_{H_i} =$	Indecision probability								
		.3			.2			.1		
		.2	.5	.8	.2	.5	.8	.2	.5	.8
2		60	19	9	108	44	30	182	108	108
3	$s$	304	74	34	567	<b>427</b>	427	*	*	*
	$m$	119	40	20	194	72	46	325	154	141
	$l$	90	31	16	158	55	32	<b>281</b>	102	83
4	$s$	294	67	29	*	366	353	*	*	*
	$m$	218	<b>55</b>	23	367	177	154	622	492	491
	$l$	102	32	15	181	59	33	324	124	101

*Note.* Let \* denote group sample sizes larger than 1,000. Small, medium, and large violations under  $H_c$  are denoted by  $s$ ,  $m$ , and  $l$ . Sample size is determined based on the allowed Indecision probability for each ordering of population means under  $H_c$ , and then averaged and presented in violation categories. Note that  $d_{H_c} = .2$  for all sample sizes.



Table 12  
*Required group sample sizes for Approach 2b using  $H_{i'}$*

$K$	$d_{H_{i'}}$	$d_{H_i} =$	Indecision probability					
			.3		.2		.1	
			.2	.5	.2	.5	.2	.5
3	$s$	.2	266	97	442	<b>190</b>	746	447
		.5	97	44	190	71	447	118
	$m$	.2	2	2	87	31	180	92
		.5	2	2	31	14	91	30
	$l$	.2	2	2	55	19	127	61
		.5	2	2	19	9	<b>61</b>	21
4	$s$	.2	600	222	990	443	*	990
		.5	222	100	443	160	990	270
	$m$	.2	2	<b>2</b>	223	86	428	227
		.5	2	2	87	37	239	68
	$l$	.2	2	2	2	2	93	38
		.5	2	2	2	2	38	16

*Note.* Let \* denote group sample sizes larger than 1,000. Small, medium, and large violations under  $H_{i'}$  are denoted by  $s$ ,  $m$ , and  $l$ .

Table 13  
Error probabilities for Approach 2b using  $H_c$

$K$	$d_{H_i} =$	Indecision probability								
		.3			.2			.1		
		.2	.5	.8	.2	.5	.8	.2	.5	.8
2	$i$	.038	.014	.008	.014	.001	.000	.005	.000	.000
	$c$	.038	.102	.132	.014	.051	.075	.005	.014	.014
	DE	.038	.058	.070	.014	.026	.037	.005	.007	.007
$s$	$i$	.008	.003	.002	.001	<b>.000</b>	.000	*	*	*
	$c$	.142	.201	.192	.080	<b>.112</b>	.112	*	*	*
	DE	.075	.102	.097	.041	<b>.056</b>	.056	*	*	*
3	$m$	$i$	.042	.010	.008	.021	.003	.000	.006	.000
	$c$	.016	.045	.066	.006	.027	.041	.001	.010	.012
	DE	.029	.028	.037	.014	.015	.021	.004	.005	.006
$l$	$i$	.056	.020	.013	.028	.005	.002	<b>.008</b>	.001	.000
	$c$	.009	.027	.047	.002	.018	.024	<b>.000</b>	.005	.008
	DE	.033	.024	.030	.015	.012	.013	<b>.004</b>	.003	.004
$s$	$i$	.015	.007	.006	*	.000	.000	*	*	*
	$c$	.364	.273	.212	*	.364	.365	*	*	*
	DE	.190	.140	.109	*	.182	.182	*	*	*
4	$m$	$i$	.031	<b>.012</b>	.011	.012	.001	.000	.003	.000
	$c$	.076	<b>.118</b>	.118	.044	.087	.093	.015	.026	.026
	DE	.053	<b>.065</b>	.065	.028	.044	.047	.009	.013	.013
$l$	$i$	.082	.035	.028	.039	.010	.004	.012	.002	.000
	$c$	.014	.038	.055	.005	.024	.039	.001	.010	.014
	DE	.048	.037	.042	.022	.017	.022	.007	.006	.007

Note. Type  $i$ , Type  $c$ , and Decision error probability are denoted by  $i$ ,  $c$ , and DE. Small, medium, and large violations under  $H_c$  are denoted by  $s$ ,  $m$ , and  $l$ . The error probabilities are determined based on the required sample size for each ordering of population means under  $H_c$ , and then averaged and presented in violation categories. Note that  $d_{H_c} = .2$  for all sample sizes. Let \* indicate that sample sizes larger than 1,000 were required to meet this level of Indecision probability, and thus no error probabilities are known.

Table 14  
Error probabilities for Approach 2b using  $H_{i'}$

$K$	$d_{H_{i'}} =$	$d_{H_i} =$	Indecision probability					
			.3		.2		.1	
			.2	.5	.2	.5	.2	.5
3	$s$	$i$	.030	.007	.015	<b>.001</b>	.005	.000
		$i'$	.030	.068	.015	<b>.044</b>	.005	.014
		DE	.030	.037	.015	<b>.023</b>	.005	.007
		$i$	.068	.028	.044	.015	.014	.005
		$i'$	.007	.028	.001	.015	.000	.005
		DE	.037	.028	.022	.015	.007	.005
	$m$	$i$	.289	.214	.053	.016	.018	.046
		$i'$	.322	.322	.053	.121	.018	.000
		DE	.305	.268	.053	.069	.018	.023
		$i$	.289	.214	.122	.054	.046	.016
		$i'$	.239	.239	.014	.058	.001	.016
		DE	.264	.227	.068	.056	.024	.016
	$l$	$i$	.316	.226	.070	.022	.022	.000
		$i'$	.328	.328	.070	.146	.022	.059
		DE	.322	.277	.070	.084	.022	.030
		$i$	.316	.226	.146	.074	<b>.059</b>	.020
		$i'$	.239	.239	.022	.074	<b>.000</b>	.020
		DE	.277	.232	.084	.074	<b>.030</b>	.020
4	$s$	$i$	.030	.007	.014	.001	*	.000
		$i'$	.030	.064	.014	.038	*	.014
		DE	.030	.035	.014	.020	*	.007
		$i$	.064	.027	.038	.014	.014	.005
		$i'$	.007	.027	.001	.014	.000	.005
		DE	.035	.027	.019	.014	.007	.005
	$m$	$i$	.255	<b>.222</b>	.043	.009	.016	.000
		$i'$	.410	<b>.410</b>	.038	.094	.013	.040
		DE	.333	<b>.316</b>	.040	.052	.015	.020
		$i$	.255	.222	.095	.044	.040	.016
		$i'$	.346	.346	.009	.038	.000	.014
		DE	.300	.284	.052	.041	.020	.015
	$l$	$i$	.320	.231	.320	.231	.037	.004
		$i'$	.383	.383	.383	.383	.038	.107
		DE	.352	.307	.352	.307	.038	.056
		$i$	.320	.231	.320	.231	.107	.036
		$i'$	.291	.291	.291	.291	.004	.037
		DE	.306	.261	.306	.261	.056	.037

Note. Type  $i$ , Type  $c$ , and Decision error probability are denoted by  $i$ ,  $c$ , and DE. Small, medium, and large violations under  $H_{i'}$  are denoted by  $s$ ,  $m$ , and  $l$ . Let \* indicate that sample sizes larger than 1,000 were required to meet this level of Indecision probability, and thus no error probabilities are known.

Table 15  
 Required group sample sizes for controlling  $BF_{ic}$  under  $H_i$  at  $B$  in Approach 3

$K$	$d_{H_i} =$	$B$								
		3			10			20		
		.2	.5	.8	.2	.5	.8	.2	.5	.8
2		24	4	2	87	15	6	136	23	9
		<i>.242</i>	<i>.238</i>	<i>.210</i>	<i>.091</i>	<i>.089</i>	<i>.091</i>	<i>.044</i>	<i>.046</i>	<i>.038</i>
3		81	13	6	301	<b>50</b>	20	513	83	33
		<i>.196</i>	<i>.210</i>	<i>.184</i>	<i>.039</i>	<b>.037</b>	<i>.037</i>	<i>.011</i>	<i>.010</i>	<i>.012</i>
4		97	16	7	352	58	23	<b>637</b>	103	40
		<i>.220</i>	<i>.224</i>	<i>.231</i>	<i>.040</i>	<i>.039</i>	<i>.040</i>	<b>.011</b>	<i>.010</i>	<i>.010</i>

Note. Entries in italics indicate  $P(BF_{ic} < 1|H_i)$ . Note that  $d_{H_c} = .2$  for all sample sizes.

Table 16  
Required group sample sizes for controlling  $BF_{ii'}$  under  $H_i$  at  $B$  in Approach 3

$K$	$d_{H_i}$	$B$					
		3		10		20	
		.2	.5	.2	.5	.2	.5
3	$s$	118	19	391	<b>63</b>	577	94
		.219	.221	.078	<b>.079</b>	.043	.040
	$m$	17	2	67	10	107	17
		.305	.331	.159	.173	.103	.107
	$l$	9	2	41	6	67	11
		.332	.332	.183	.195	.123	.126
4	$s$	274	45	884	144	*	212
		.218	.213	.079	.072	—	.040
	$m$	56	2	187	30	<b>283</b>	45
		.280	.345	.137	.138	<b>.091</b>	.092
	$l$	2	2	24	2	<b>38</b>	2
		.394	.296	.230	.296	<b>.180</b>	.296

Note. Entries in italics indicate  $P(BF_{ic} < 1|H_i)$ . Let \* denote sample sizes larger than 1,000, and — denote the absence of  $P(BF_{ic} < 1|H_i)$  in this situation.

Table 17  
*Required group sample sizes for controlling  $BF_{ic}$  under  $H_c$  at  $1/B$  in Approach 3*

$K$		$B$		
		3	10	20
2		24	87	136
		.242	.091	.044
		444	<b>839</b>	*
	$s$	.281	<b>.134</b>	—
3	$m$	21	110	165
		.247	.095	.061
	$l$	<b>2</b>	58	101
		<b>.303</b>	.091	.049
	$s$	2	*	*
		.303	—	—
4	$m$	2	370	492
		.272	.146	.099
	$l$	2	52	107
		.228	.146	.071

*Note.* Entries in italics indicate  $P(BF_{ic} > 1|H_c)$ . Let \* denote sample sizes larger than 1,000, and — denote the absence of  $P(BF_{ic} > 1|H_c)$  in this situation. Note that  $d_{H_c} = .2$  for all sample sizes.

Table 18  
 Required group sample sizes for controlling  $BF_{ii'}$  under  $H_{i'}$  at  $1/B$  in Approach 3

$K$		3	10	20	
3	$s$	.2	118	<b>391</b>	577
			<i>.219</i>	<i>.078</i>	<i>.043</i>
		.5	19	63	94
			<i>.221</i>	<i>.079</i>	<i>.040</i>
	$m$	.2	17	67	105
			<i>.303</i>	<i>.158</i>	<i>.103</i>
		.5	3	11	17
			<i>.312</i>	<i>.161</i>	<i>.100</i>
	$l$	.2	<b>9</b>	41	67
			<b>.332</b>	<i>.183</i>	<i>.123</i>
		.5	2	6	11
			<i>.334</i>	<i>.196</i>	<i>.126</i>
4	$s$	.2	274	884	*
			<i>.218</i>	<i>.079</i>	—
		.5	45	144	212
			<i>.214</i>	<i>.073</i>	<i>.040</i>
	$m$	.2	50	179	271
			<i>.282</i>	<i>.131</i>	<i>.085</i>
		.5	9	30	43
			<i>.287</i>	<i>.125</i>	<i>.082</i>
	$l$	.2	4	24	40
			<i>.385</i>	<i>.231</i>	<i>.175</i>
		.5	2	4	6
			<i>.359</i>	<i>.242</i>	<i>.196</i>

Note. Entries in italics indicate  $P(BF_{ic} > 1|H_c)$ . Let \* denote sample sizes larger than 1,000, and — denote the absence of  $P(BF_{ic} > 1|H_c)$  in this situation.

## A Appendix: Numerical characteristics of tables

This appendix illustrates some numerical characteristics of Tables 5–18 by means of examples.

First, in all tables the required sample size increases if the error probability or In-decision probability decreases, or if  $B$  increases. Put differently, the more certainty is required for the conclusion, the larger the sample size should be. If the violation size under  $H_{i'}$  increases, the required sample size decreases. Hypotheses with larger violations are more distinctly different from  $H_i$ : datasets generated under  $H_i$  will less often result in a decision in favour of  $H_{i'}$ , and vice versa, compared to small violations.

Second, if  $K$  increases, a larger sample size is required. If  $K$  increases, but  $d_{H_i}$  is constant, the differences between pair of means decreases. For example, if  $d_{H_i} = .5$ , the difference between each pair of means is .5 for  $K = 2$ , .25 for  $K = 3$ , and .167 for  $K = 4$ . If differences between means are smaller, it is more likely that the means of a sample will not adhere to the population from which they were sampled, thus, a larger sample size is required.

Furthermore, in three situations, some symmetry is visible in the tables. First, if  $d_{H_i}$  and  $d_{H_c}$  are equal with  $K = 2$ ,  $H_i$  and  $H_c$  are exchangeable. Therefore, the Type  $i$  and Type  $c$  error probabilities are equal. Since the Decision error probability is their average, it holds that this is equal to both the Type  $i$  and Type  $c$  error probability. Thus, the required sample size is independent of the type of error probability controlled. For example, as can be seen in Table 5, for  $K = 2$ ,  $d_{H_i} = .2$ , and a critical value for the error probabilities of .05, the group sample size is 132, whether the Decision error, Type  $i$ , or Type  $c$  error probability is controlled.

Second, if  $d_{H_i}$  and  $d_{H_{i'}}$  are equal,  $H_i$  and  $H_{i'}$  are exchangeable as well, and thus the Type  $i$  and Type  $i'$  error probabilities are equal. Thus, the sample size is independent of the type of error controlled. As can be seen in Table 6, for  $K = 3$ ,  $d_{H_i} = d_{H_{i'}} = .5$ , and  $H_{i'}$  with a large violation size, and a critical value for the error probability of .05, the sample size is 22, whether the Decision error, Type  $i$ , or Type  $c$  error probability is controlled.

Third,  $d_{H_i}$  and  $d_{H_{i'}}$  can be unequal in two ways:  $d_{H_i} = .2$  and  $d_{H_{i'}} = .5$ , or  $d_{H_i} = .5$  and  $d_{H_{i'}} = .2$ . The Type  $i$  error probability for  $d_{H_i} = .2$  and  $d_{H_{i'}} = .5$  will be the same as the Type  $i'$  error probability for  $d_{H_i} = .5$  and  $d_{H_{i'}} = .2$ , and vice versa. Thus, the Decision error probability will be exchangeable in these situations. Therefore, the required sample size to control the Decision error probability will be the same whether the first or the second set of effect sizes is used. As can be seen in Table 6, for  $K = 4$ ,  $H_{i'}$  with a small violation, and a critical value for the Decision error probability of .1, for  $d_{H_i} = .5$  and  $d_{H_{i'}} = .2$  or  $d_{H_i} = .2$  and  $d_{H_{i'}} = .5$ , the sample size is 338.

Note that examples of the three situations described above exist where the sample sizes are not exactly equal, but about equal. This is caused by the sampling variation of the simulation procedure presented in Section 4.4.

Contrary to expectations, the required sample size when comparing  $H_i$  with  $H_{i'}$  is



not always smaller than when comparing  $H_i$  with  $H_c$ . As an example, a repetition of Example 1.3 from Section 5.1 follows:

*Example 1.3* Suppose a researcher wants to evaluate  $H_i$  with  $H_c$ , with  $K = 4$ . The researcher wants to control the Type  $i$  error probability at .025. He specifies  $d_{H_i} = .2$ ,  $d_{H_c} = .2$ . As can be seen in Table 5, the required sample size is 443. Suppose this researcher did not consider  $H_c$ , but  $H_{i'}$ . As can be seen in Table 6, the required sample size is larger than 1,000 if  $H_{i'}$  was specified with a small violation of  $H_i$ , the sample size is 575 with a medium violation, and 169 with a large violation.

In this example, the required sample size to control the Type  $i$  error comparing  $H_i$  with  $H_c$  is smaller than the sample size when comparing  $H_i$  with  $H_{i'}$ , only for medium and small violations of  $H_{i'}$ , but not for large violations of  $H_{i'}$ . This can be explained best by means of an example: Let us consider  $K = 4$ , such that  $c_i = c_{i'} = 1/24$ , and  $c_c = 23/24$ , then,

$$BF_{ic} = \frac{f_i/c_i}{f_c/c_c} = \frac{f_i/\frac{1}{24}}{f_c/\frac{23}{24}} = \frac{f_i}{f_c/23} = f_i \cdot 23 \frac{1}{f_c}$$

$$BF_{ii'} = \frac{f_i/c_i}{f_{i'}/c_{i'}} = \frac{f_i/\frac{1}{24}}{f_{i'}/\frac{1}{24}} = \frac{f_i}{f_{i'}} = f_i \cdot \frac{1}{f_{i'}}$$

Thus, if the fit of the data to  $H_c$  is five times larger than the fit of the data to  $H_{i'}$ ,  $BF_{ic}$  and  $BF_{ii'}$  will be equal. If the fit of the data to  $H_c$  is less than five times the fit to  $H_{i'}$ ,  $BF_{ic}$  will be larger than  $BF_{ii'}$ , and if it is more than five times the fit to  $H_{i'}$ ,  $BF_{ic}$  will be smaller than  $BF_{ii'}$ . If Bayes factors based on population in which  $H_i$  is true are larger, the Type  $i$  error probability becomes smaller.

In Example 1.3, only the Type  $i$  error is considered, which means that only data generated under  $H_i$  is considered. Thus,  $f_i$  will be the same for  $BF_{ic}$  and for  $BF_{ii'}$ . However, the fit to  $H_c$  or to  $H_{i'}$  is also taken into account in the computation of the Bayes factor. Any part of the posterior distribution that does not fit to  $H_i$ , will fit to  $H_c$ . Since  $H_{i'}$  is a subset of  $H_c$ , the fit to  $H_{i'}$  will always be smaller than that to  $H_c$ .

More often than not, the fit of data generated under  $H_i$  will be larger to  $H_{i'_{\text{small}}}$ , than to  $H_{i'_{\text{large}}}$ . If data are generated based on a population in which  $H_i$  holds, it is more likely to obtain a sample that adheres to  $H_{i'_{\text{small}}}$  than to  $H_{i'_{\text{large}}}$ . In general, using data simulated from a population in which  $H_i$  is true, Bayes factors concerning  $H_{i'_{\text{large}}}$  will be larger than those concerning  $H_{i'_{\text{small}}}$ , and thus the Type  $i$  error probability will be smaller for  $H_{i'_{\text{large}}}$  than for  $H_{i'_{\text{small}}}$ .

This explains the differences in sample size for the different violation sizes under  $H_{i'}$ . Applying this to Example 1.3, it appears that for a large violation under  $H_{i'}$ , the fit to  $H_c$  is more than 23 times larger than to  $H_{i'}$ . thus the required sample size for a large violation is smaller than that for  $H_c$ .