Bayes Factor Design Analysis: Planning for Compelling Evidence

Felix D. Schönbrodt
Ludwig-Maximilians-Universität München

Eric-Jan Wagenmakers
University of Amsterdam

A sizeable literature exists on the use of frequentist power analysis in the null-hypothesis significance testing (NHST) paradigm to facilitate the design of informative experiments. In contrast, there is almost no literature that discusses the design of experiments when Bayes factors (BFs) are used as a measure of evidence. Here we explore Bayes Factor Design Analysis (BFDA) as a useful tool to design studies for maximum efficiency and informativeness. We elaborate on three possible BF designs, (a) a fixed-*n* design, (b) an open-ended Sequential Bayes Factor (SBF) design, where researchers can test after each participant and can stop data collection whenever there is strong evidence for either $\mathcal{H}_1$ or $\mathcal{H}_0$, and (c) a modified SBF design that defines a maximal sample size where data collection is stopped regardless of the current state of evidence. We demonstrate how the properties of each design (i.e., expected strength of evidence, expected sample size, expected probability of misleading evidence, expected probability of weak evidence) can be evaluated using Monte Carlo simulations and equip researchers with the necessary information to compute their own Bayesian design analyses.

Manuscript accepted for publication in *Psychonomic Bulletin & Review*.

Keywords: Bayes factor, power analysis, design analysis, design planning, sequential testing

> "The following rule of experimentation is therefore suggested: perform that experiment for which the expected gain in information is the greatest, and continue experimentation until a preassigned amount of information has been attained" (Lindley, 1956, p. 987)

We aim to explore *Bayes Factor Design Analysis* (BFDA) as a useful tool to design studies for maximum efficiency and informativeness. In the classical frequentist framework, statistical power refers to the long-term probability (across multiple hypothetical studies) of obtaining a significant *p*-value in case an effect of a certain size exists (Cohen, 1988). Classical power analysis is a special case of the broader class of *design analysis*, which uses prior guesses of effect sizes and other parameters in order to compute distributions of any study outcome (Gelman & Carlin, 2014).[1] The general principle is to assume a certain state of reality, most importantly the expected true effect size, and tune the settings of a research design in a way such that certain desirable outcomes are likely to occur. For example, in frequentist power analysis, the property "sample size" of a design can be tuned such that, say, 80% of all studies would yield a p-value $< .05$ if an effect of a certain size exists.

The framework of design analysis is general and can be used both for Bayesian and non-Bayesian designs, and it can be applied to any study outcome of interest. For example, in designs reporting Bayes factors a researcher can plan sample size such that, say, 80% of all studies result in a compelling Bayes factor, for instance $BF_{10} > 10$ (De Santis, 2004; Weiss, 1997). One can also determine the sample size such that, with a desired probability of occurrence, a highest density interval for a parameter excludes zero, or a particular parameter is estimated with a predefined precision (Gelman & Tuerlinckx, 2000; Kruschke, 2014). Hence, the concept of prospective design analysis, which refers to design planning before data are collected, is not limited to null-hypothesis significance testing (NHST), and our paper applies the concept to studies that use Bayes factors (BFs) as an index of evidence.

The first part of this article provides a short introduction

Felix D. Schönbrodt, Department of Psychology, Ludwig-Maximilians-Universität München, Germany. Eric-Jan Wagenmakers, University of Amsterdam. Reproducible code and figures are available at https://osf.io/qny5x. Acknowledgements. This research was supported in part by grant 283876 "Bayes or Bust!" awarded by the European Research Council. Correspondence concerning this article should be addressed to Felix Schönbrodt, Leopoldstr. 13, 80802 München, Germany. Email: felix@nicebread.de. Phone: +49 89 2180 5217. Fax: +49 89 2180 99 5214.

[1] Other authors have used "power analysis" as a generic term for the "probability of achieving a research goal" (e.g. Kruschke, 2010, p. 1). In line with Gelman and Carlin (2014), we prefer the more general term "design analysis" and reserve "power analysis" for the special case where a design analysis aims to ensure a minimum rate of true positive outcomes in a hypothesis test (i.e., prob(strong $\mathcal{H}_1$ evidence $| \mathcal{H}_1$), which is the classical meaning of statistical power.

to BFs as a measure of evidence for a hypothesis (relative to an alternative hypothesis). The second part describes how compelling evidence is a necessary ingredient for strong inference, which has been argued to be the fastest way to increase knowledge (Platt, 1964). The third part of this article elaborates on how to apply the idea of design analysis to research designs with BFs. The fourth part introduces three BF designs, (a) a fixed-n design, (b) an open-ended Sequential Bayes Factor (SBF) design, where researchers can test after each participant and can stop data collection when there is strong evidence for either $\mathcal{H}_1$ or $\mathcal{H}_0$, and (c) a modified SBF design that defines a maximal sample size where data collection is stopped in any case. We demonstrate how to use Monte Carlo simulations and graphical summaries to assess the properties of each design and how to plan for compelling evidence. Finally, we discuss the approach in terms of possible extensions, the issue of (un)biased effect size estimates in sequential designs, and practical considerations.

Bayes Factors as an Index of Evidence

The Bayes factor is "fundamental to the Bayesian comparison of alternative statistical models" (O'Hagan & Forster, 2004, p. 55) and it represents "the standard Bayesian solution to the hypothesis testing and model selection problems" (Lewis & Raftery, 1997, p. 648) and "the primary tool used in Bayesian inference for hypothesis testing and model selection" (Berger, 2006, p. 378). Here we briefly describe the Bayes factor as it applies to the standard scenario where a precise, point-null hypothesis $\mathcal{H}_0$ is compared to a composite alternative hypothesis $\mathcal{H}_1$. Under a composite hypothesis, the parameter of interest is not restricted to a particular fixed value (Jeffreys, 1961). In the case of a t-test, for instance, the null hypothesis specifies the absence of an effect, that is, $\mathcal{H}_0 : \delta = 0$, whereas the composite alternative hypothesis allows effect size to take on nonzero values.

In order to gauge the support that the data provide for $\mathcal{H}_0$ versus $\mathcal{H}_1$, the Bayes factor hypothesis test requires that both models make predictions. This, in turn, requires that the expectations under $\mathcal{H}_1$ are made explicit by assigning effect size δ a prior distribution, for instance a normal distribution centered on zero with a standard deviation of 1, $\mathcal{H}_1 : \delta \sim \mathcal{N}(0, 1)$.

After both models have been specified so that they make predictions, the observed data can be used to assess each models' predictive adequacy (Morey, Romeijn, & Rouder, 2016; Wagenmakers, Grünwald, & Steyvers, 2006; Wagenmakers, Morey, & Lee, 2016). The ratio of predictive adequacies –the Bayes factor– represents the extent to which the data update the relative plausibility of the competing hypotheses, that is:

$$\underbrace{\frac{p(\mathcal{H}_0 \mid \text{data})}{p(\mathcal{H}_1 \mid \text{data})}}_{\substack{\text{Posterior plausibility}\\ \text{about hypotheses}}} = \underbrace{\frac{p(\mathcal{H}_0)}{p(\mathcal{H}_1)}}_{\substack{\text{Prior plausibility}\\ \text{about hypotheses}}} \times \underbrace{\frac{p(\text{data} \mid \mathcal{H}_0)}{p(\text{data} \mid \mathcal{H}_1)}}_{\substack{\text{Bayes factor =}\\ \text{Predictive updating factor}}} \quad (1)$$

In this equation, the relative prior plausibility of the competing hypotheses is adjusted in light of predictive performance for observed data, and this then yields the relative posterior plausibility. Although the assessment of prior plausibility may be informative and important (e.g., Dreber et al., 2015), the inherently subjective nature of this component has caused many Bayesian statisticians to focus on the Bayes factor –the predictive updating factor– as the metric of interest (Hoijtink, Klugkist, & Boelen, 2008; Jeffreys, 1961; Kass & Raftery, 1995; Ly, Verhagen, & Wagenmakers, 2016; Mulder & Wagenmakers, 2016; Rouder, Morey, Speckman, & Province, 2012; Rouder, Speckman, Sun, Morey, & Iverson, 2009).

Depending on the order of numerator and denominator in the ratio, the Bayes factor is either denoted as BF_{01} ("$\mathcal{H}_0$ over $\mathcal{H}_1$", as in Eq. (1)) or as its inverse BF_{10} ("$\mathcal{H}_1$ over $\mathcal{H}_0$"). When the Bayes factor BF_{01} equals 5, this indicates that the data are five times more likely under $\mathcal{H}_0$ than under $\mathcal{H}_1$, meaning that $\mathcal{H}_0$ has issued a better probabilistic prediction for the observed data than did $\mathcal{H}_1$. In contrast, when BF_{01} equals 0.25 the data support $\mathcal{H}_1$ over $\mathcal{H}_0$. Specifically, the data are $1/\text{BF}_{01} = \text{BF}_{10} = 4$ times more likely under $\mathcal{H}_1$ than under $\mathcal{H}_0$.

The Bayes factor offers several advantages for the practical researcher (Wagenmakers et al., 2016). First, the Bayes factor quantifies evidence, both for $\mathcal{H}_1$ but also for $\mathcal{H}_0$; second, its predictive underpinnings entail that neither $\mathcal{H}_0$ nor $\mathcal{H}_1$ need be "true" for the Bayes factor to be useful (but see van Erven, Grünwald, & de Rooij, 2012); third, the Bayes factor does not force an all-or-none decision, but instead coherently reallocates belief on a continuous scale; fourth, the Bayes factor distinguishes between absence of evidence and evidence of absence (e.g., Dienes, 2014, 2016); fifth, the Bayes factor does not require adjustment for sampling plans (i.e., the Stopping Rule Principle; Bayarri, Benjamin, Berger, & Sellke, 2016, Berger & Wolpert, 1988, Rouder, 2014). A practical corollary is that, in contrast to p-values, Bayes factors retain their meaning in situations common in ecology and astronomy, where nature provides data over time and sampling plans do not exist (Wagenmakers et al., 2016).

Although Bayes factors are defined on a continuous scale, several researchers have proposed to subdivide the scale in discrete evidential categories (Jeffreys, 1961; Kass & Raftery, 1995; Lee & Wagenmakers, 2013). The scheme originally proposed by Jeffreys is shown in Table 1. The evidential categories serve as a rough heuristic whose main goal is to prevent researchers from overinterpreting the evidence in the data. In addition –as we will demonstrate below– the

categories permit a concise summary of the results from our simulation studies.

Table 1
A rough heuristic classification scheme for the interpretation of Bayes factors BF_{10} (Lee and Wagenmakers 2013; adjusted from Jeffreys 1961).

Bayes factor	Evidence category
> 100	Extreme evidence for $\mathcal{H}_1$
30 - 100	Very strong evidence for $\mathcal{H}_1$
10 - 30	Strong evidence for $\mathcal{H}_1$
3 - 10	Moderate evidence for $\mathcal{H}_1$
1 - 3	Anecdotal evidence for $\mathcal{H}_1$
1	No evidence
1/3 - 1	Anecdotal evidence for $\mathcal{H}_0$
1/10 - 1/3	Moderate evidence for $\mathcal{H}_0$
1/30 - 1/10	Strong evidence for $\mathcal{H}_0$
1/100 - 1/30	Very strong evidence for $\mathcal{H}_0$
< 1/100	Extreme evidence for $\mathcal{H}_0$

The Purpose of Design Analyses: Planning for Compelling Evidence

In the planning phase of an experiment, the purpose of a prospective design analysis is to facilitate the design of a study that ensures a sufficiently high probability of detecting an effect if it exists. Executed correctly, this is a crucial ingredient to *strong inference* (Platt, 1964), which includes "[d]evising a crucial experiment [...], with alternative possible outcomes, each of which will, as nearly as possible, exclude one or more of the hypotheses" (p. 347). In other words, a study design with strong inferential properties is likely to provide compelling evidence, either for one hypothesis or for the other. Such a study generally does not leave researchers in a state of inference that is inconclusive.

When a study is underpowered, in contrast, it most likely provides only weak inference. Within the framework of frequentist statistics, underpowered studies result in *p*-values that are relatively nondiagnostic. Specifically, underpowered studies inflate both false-negative and false-positive results (Button et al., 2013; Dreber et al., 2015; Ioannidis, 2005; Lakens & Evers, 2014), wasting valuable resources such as the time and effort of participants, the lives of animals, and scientific funding provided by society. Consequently, research unlikely to produce diagnostic outcomes is inefficient and can even be considered unethical (Emanuel, Wendler, & Grady, 2000, Halpern, Karlawish, & Berlin, 2002; but see Bacchetti, Wolf, Segal, & McCulloch, 2005).

To summarize, the primary purpose of a prospective design analysis is to assist in the design of studies that increase the probability of obtaining compelling evidence, a necessary requirement for strong inference.

Design Analysis for Bayes Factor Designs

We apply design analysis to studies that report the Bayes factor as a measure of evidence. Note, first, that we seek to evaluate the operational characteristics of a Bayesian research design *before* the data are collected (i.e., a prospective design analysis). Therefore, our work centers on design, not on inference; once specific data have been collected, pre-data design analyses are inferentially irrelevant, at least from a Bayesian perspective (Bayarri et al., 2016; Wagenmakers et al., 2014). Second, our focus is on the Bayes factor as a measure of evidence, and we expressly ignore both prior model probabilities and utilities (Berger, 1985; Lindley, 1997; Taroni, Bozza, Biedermann, Garbolino, & Aitken, 2010), two elements that are essential for decision making yet orthogonal to the quantification of evidence provided by the observed data. Thus, we consider scenarios where "the object of experimentation is not to reach decisions but rather to gain knowledge about the world" (Lindley, 1956, p. 986).

Target Outcome of a Bayes Factor Design Analysis: Strong Evidence and No Misleading Evidence

In the context of evaluating the empirical support for and against a null hypothesis, Bayes factors quantify the strength of evidence for that null hypothesis $\mathcal{H}_0$ relative to the alternative hypothesis $\mathcal{H}_1$. To facilitate strong inference, we wish to design studies such that they are likely to result in compelling Bayes factors in favor of the true hypothesis – thus, the informativeness of a design may be quantified by the expected Bayes factor (Cavagnaro, Myung, Pitt, & Kujala, 2009; Good, 1979; Lindley, 1956), or an entire distribution of Bayes factors.

Prior to the experiment, one may expect that in the majority of data sets that may be obtained the Bayes factor will point towards the correct hypothesis. However, for particular data sets sampling variability may result in a misleading Bayes factor, that is, a Bayes factor that points towards the incorrect hypothesis. For example, even when $\mathcal{H}_0$ holds in the population, a random sample can show strong evidence in favor of $\mathcal{H}_1$, just by sampling fluctuations. We term this situation *false positive evidence* (FPE). If, in contrast, the data set shows strong evidence for $\mathcal{H}_0$, although in reality $\mathcal{H}_1$ is correct, we term this *false negative evidence* (FNE). In general terms, misleading evidence is defined as a situation where the data show strong evidence in favor of the incorrect hypothesis (Royall, 2000).

Research designs differ with respect to their probability of generating misleading evidence. The probability of yielding misleading evidence is a pre-data concept that should not be confused with a related but different post-data concept,

namely the probability that a given evidence in a particular data set is misleading (Blume, 2002).

The expected strength of evidence (i.e., the expected BF) and the probability of misleading evidence are conceptually distinct, but practically tightly related properties of a research design (Royall, 2000), as in general higher evidential thresholds will lead to lower rates of misleading evidence (Blume, 2008; Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, 2015). To summarize, the joint goal of a prospective design analysis should be a high probability of obtaining strong evidence and a low probability of obtaining misleading evidence, which usually go together.

Dealing With Uncertainty in Expected Effect Size

Power in a classical power analysis is a *conditional power*, because the computed power is conditional on the assumed true (or minimally interesting) effect size. One difficulty is to commit to a point estimate of that parameter when there is considerable uncertainty about it. This uncertainty could be dealt with by computing the necessary sample size for a *set* of plausible fixed parameter values. For example, previous experiments may suggest that the true effect size is around 0.5, but a researcher feels that the true effect could as well be 0.3 or 0.7, and computes the necessary sample sizes for these effect size guesses as well. Such a sensitivity analysis gives an idea about the variability of resulting sample sizes.

A problem of this approach, however, is that there is no principled way of choosing an appropriate sample size from this set: Should the researcher aim for the conservative estimate, which would be highly inefficient in case the true effect is larger? Or should she aim for the optimistic estimate, which would lead to a low actual power if the true effect size is at the lower end of plausible values?

Prior effect size distributions quantify uncertainty. Extending the procedure of a sensitivity analysis, however, one can compute the probability of achieving a research goal averaged across all possible effect sizes. For this purpose, one has to define prior plausibilities of the effect sizes, compute the distribution of target outcomes for each effect size, and then obtain a weighted average. This averaged probability of success has been called "assurance" (O'Hagan, Stevens, & Campbell, 2005) or "expected Bayesian power" (Spiegelhalter, Abrams, & Myles, 2004), and is the expected probability of success with respect to the prior.[2]

In the above example, not all of the three assumed effect sizes (i.e, 0.3, 0.5, and 0.7) might be equally plausible. For example, one could construct a prior effect size distribution under $\mathcal{H}_1$ that describes the plausibility for each choice (and all effect sizes in between) as a normal distribution centered around the most plausible value of 0.5 with a standard deviation of 0.1: $\delta \sim \mathcal{N}(0.5, \sigma = 0.1)$, see Figure 1.

Garthwaite, Kadane, and O'Hagan (2005) give advice on how to elicit a prior distribution from experts. These procedures help an expert to formulate his or her substantive knowledge in probabilistic form, which in turn can be used for Bayesian computations. Such an elicitation typically includes several steps, for example asking experts about the most plausible value (i.e., about the mode of the prior), or asking about the quantiles, such as 'Please make a guess about a very high value, such that you feel there is only a 5% probability the true value would exceed your guess'.

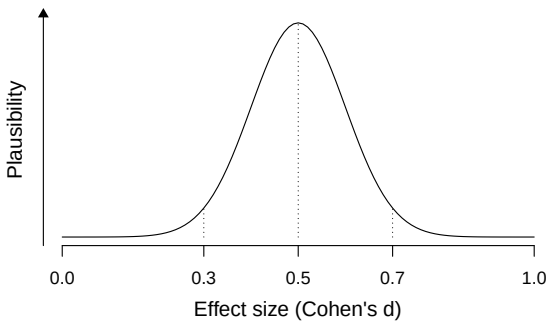

Figure 1. A hypothetical prior distribution expressing the uncertainty about the true effect size. Figure available at https://osf.io/qny5x/, under a CC-BY4.0 license.

Morris, Oakley, and Crowe (2014) provide an online tool that can help to fit an appropriate distribution to an experts' input[3].

Design priors vs. analysis priors. Two types of priors can be differentiated (O'Hagan & Stevens, 2001 May-Jun; Walley, Smith, Gale, & Woodward, 2015). *Design priors* are used before data collection to quantify prior beliefs about the true state of nature. These design priors are used to do design analyses and in general to assist experimental design. *Analysis priors*, in contrast, are used for Bayesian statistical analysis after the data are in.

At first glance it might appear straightforward to use the same priors for design planning and for data analysis. Both types of priors, however, can serve different goals. The de-

[2] It is possible to construct an unconditional effect size prior that describes the plausibility of effect sizes both under $\mathcal{H}_1$ and $\mathcal{H}_0$, for example by defining a prior effect size distribution that assigns considerable probability to values around zero and the opposite direction, or by using a mixture distribution that has some mass around zero, and some mass around a non-zero effect size (Muirhead & Şoaita, 2013). Here, in contrast, we prefer to construct a conditional effect size prior under $\mathcal{H}_1$ and to contrast it with a point $\mathcal{H}_0$ that has all probability mass on zero. Hence, the result of our design analysis is a conditional average probability of success under $\mathcal{H}_1$, which Eaton, Muirhead, and Soaita (2013) consider to be the most plausible average probability for sample size planning.

[3] http://optics.eee.nottingham.ac.uk/match/uncertainty.php

sign prior is used to tune the design before data collection to make compelling evidence likely and to avoid misleading evidence. The target audience for a design analysis is mainly the researcher him- or herself, who wants to design the most informed study. Hence, design priors should be based on the researcher's experience and can contain a lot of existing prior information and experience to aid an optimal planning of the study's design. Relying on a non-central, highly informative prior (in the extreme case, a point effect size guess as in classical power analysis) can result in a highly efficient design (i.e., with a just large-enough sample size) if the real effect size is close to that guess. On the other hand, it bears the risk to end up with inconclusive evidence if the true effect is actually smaller. A less informative design prior, in contrast, will typically lead to larger planned sample sizes, as more plausibility is assigned to smaller effect sizes.[4]

This increases the chances of compelling evidence in the actual data analysis, but can be inefficient compared to a design that uses a more precise (and valid) effect size guess. Researchers may balance that trade-off based on their subjective certainty about plausible effect sizes, utilities about successful or failed studies, or budget constraints. Whenever prospective design analyses are used to motivate sample size costs in grant applications, the design priors should be convincing to the funder and the grant reviewers.

The analysis priors that are used to compute the BF, in contrast, should be convincing to a skeptical target audience, and therefore often are less informative than the design priors. In the examples of this paper, we will use an informed, non-central prior distribution for the planning stage, but a default effect size prior (which is less informative) for data analysis.

Three Exemplary Designs for a Bayes Factor Design Analysis

In the next sections, we will demonstrate how to conduct a Bayes Factor Design Analyses. We consider three design perspectives:

1. *Fixed*-n *design*: In this design, a sample of fixed size is collected and one data analysis is performed at the end. From this perspective, one can ask the following design-related questions: Given a fixed sample size and the expected effect size – what BFs can be expected? What sample size do I need to have at least a 90% probability of obtaining a BF_{10} of, say, 6 or greater? What is the probability of obtaining misleading evidence?

2. *Open-ended sequential designs*: Here participants are added to a growing sample and BFs are computed until a desired level of evidence is reached (Schönbrodt et al., 2015). As long as researchers do not run out of participants, time, or money, this approach eliminates the possibility of ending up with weak evidence. With this design, one can ask the following design-related questions: Given the desired level of evidence and the expected effect size – what distribution of sample sizes can be expected? What is the probability of obtaining misleading evidence?

3. *Sequential designs with maximal* n: In this modification of the open-ended SBF design, participants are added until (a) a desired level of evidence is obtained, or (b) a maximum number of participants has been reached. If sampling is stopped because of (b), the evidence will not be as strong as desired initially, but the direction and the strength of the BF can still be interpreted. With this design, one can ask the following design-related questions: Given the desired level of evidence, the expected effect size, and the maximum sample size – what distribution of sample sizes can be expected? How many studies can be expected to stop because of crossing the evidential threshold, and how many because $n_{\max}$ has been reached? What is the probability of obtaining misleading evidence?

As most design planning concerns directional hypotheses, we will focus on these in this paper. Furthermore, in our examples we use the JZS default Bayes factor for a two group t-test provided in the *BayesFactor* package (Morey & Rouder, 2015) for the R Environment for Statistical Computing (R Core Team, 2014) and in JASP (JASP Team, 2016). The JZS Bayes factor assumes that effect sizes under $\mathcal{H}_1$ (expressed as Cohen's d) follow a central Cauchy distribution (Rouder et al., 2009). The Cauchy distribution with a scale parameter of 1 equals a t distribution with one degree of freedom. This prior has several convenient properties and can be used as a default choice when no specific information about the expected effects sizes is available. The width of the Cauchy distribution can be tuned using the scale parameter, which corresponds to smaller or larger plausible effect sizes. In our examples below, we use a default scale parameter of $\sqrt{2}/2$. This corresponds to the prior expectation that 50% of probability mass is placed on effect sizes that have an (absolute) size smaller than $\sqrt{2}/2$, and 50% larger than $\sqrt{2}/2$. Note that all computations and procedures outlined here are not restricted to these specific choices and can be easily general-

[4] In Bayesian parameter estimation so called uninformative priors are quite common. A very wide prior, such as a half-normal distribution with mean=0 and SD=10, however, should not be used for design analysis, as too much probability mass is placed upon unrealistically large effect sizes. Such a design analysis will yield planned sample sizes that are usually too small, and consequently the actual data analysis will most likely be uninformative. As any design choice involves the fundamental trade-off between expected strength of evidence and efficiency, there exists no "uninformative" design prior in prospective design analysis.

ized to undirected tests and all other flavors of Bayes factors as well (Dienes, 2014).

Fixed-*n* Design

With a pre-determined fixed sample size, two related questions can be asked in a design analysis: (a) What is the expected distribution of obtained evidence? (b) What is the probability of obtaining misleading evidence? (c) Sample size determination: What is the necessary sample size that compelling evidence can be expected with sufficiently high probability?

Monte Carlo simulations can be used to answer these questions easily. In our example, we focus on a test for the difference between two population means (i.e., a Bayesian *t*-test; Rouder et al., 2009). For didactic purposes, we demonstrate this design analysis with a fixed expected effect size (i.e., without a prior distribution). This way the design analysis is analogous to a classical power analysis in the NHST paradigm, that also assumes a fixed effect size under $\mathcal{H}_1$.

The recipe for our Monte Carlo simulations is as follows (see also Kruschke, 2014):

1. Define a population that reflects the expected effect size under $\mathcal{H}_1$ and, if prior information is available, other properties of the real data (e.g., specific distributional properties). In the example given below, we used two populations with normal distributions and a fixed standardized mean difference of $\delta = 0.5$.

2. Draw a random sample of size n_{fixed} from the populations (all n refer to sample size in each group).

3. Compute the BF for that simulated data using the analysis prior that will also be used in the actual data analysis and save the result. In the example given below, we analyzed simulated data with a Cauchy prior (scale parameter = $\sqrt{2}/2$).

4. Repeat steps 2 and 3, say, 10,000 times.

5. In order to compute the probability of false-positive evidence, the same simulation must be done under the $\mathcal{H}_0$ (i.e., two populations that have no mean difference).

Researchers do not know in advance whether and to what extent the data will support $\mathcal{H}_1$ or $\mathcal{H}_0$; therefore, all simulations must be carried out both under $\mathcal{H}_1$ and $\mathcal{H}_0$ (see step 5). Figure 2 provides a flow chart of the simulations that comprise a Bayes factor design analysis. For standard designs, readers can conduct their own design analyses simulations using the R package BFDA (Bayes factor design analysis; Schönbrodt, 2016, see https://github.com/nicebread/BFDA).[5]

The proposed simulations provide a distribution of obtained BFs under $\mathcal{H}_1$, and another distribution under $\mathcal{H}_0$. For these distributions, one can set several thresholds and retrieve the probability that a random study will provide a BF in a certain evidential category. For example, one can set a single threshold at $BF_{10} = 1$ and compute the probability of obtaining a BF with the wrong direction. Or, one can aim for more compelling evidence and set thresholds at $BF_{10} = 6$ and $BF_{10} = 1/6$. This means evidence is deemed inconclusive when $1/6 < BF_{10} < 6$. Furthermore, one can define asymmetric thresholds under $\mathcal{H}_0$ and $\mathcal{H}_1$. Depending on the analysis prior in the computation of the BF, it can be expensive and time-consuming to gather strong evidence for $\mathcal{H}_0$. In these cases one can relax the requirements for strong $\mathcal{H}_0$ support and still aim for strong $\mathcal{H}_1$ support, for example by using thresholds 1/6 and 20 (Weiss, 1997).

Expected distribution of BFs and rates of misleading evidence. Figure 3 compares the BF_{10} distribution that can be expected under $\mathcal{H}_1$ (top row) and under $\mathcal{H}_0$ (bottom row). The simulations were conducted with two fixed sample sizes: $n = 20$ (left column) and $n = 100$ (right column). Evidence thresholds were defined at 1/6 and 6. If an effect of $\delta = 0.5$ exists and studies with $n = 20$ are conducted, 0.3% of all simulated studies point towards the (wrong) $\mathcal{H}_0$ ($BF < 1/6$). This is the rate of false negative evidence, and it is visualized as the dark grey area in the top density of Figure 3A. Conversely, 21.1% of studies show $\mathcal{H}_1$ support ($BF_{10} > 6$; light gray area in the top density), which is the probability of true positive results. The remaining 78.5% of studies yield inconclusive evidence ($1/6 < BF_{10} < 6$; medium grey area in the top density).

If, however, no effect exists (see bottom density of Figure 3A), 0.9% of all studies will yield false-positive evidence ($BF_{10} > 6$), and 13.7% of all studies correctly support $\mathcal{H}_0$ with the desired strength of evidence ($BF_{10} < 1/6$). A large majority of studies (85.5%) remain inconclusive under $\mathcal{H}_0$ with respect to that threshold. Hence, a design with that fixed sample size has a high probability of being uninformative under $\mathcal{H}_0$.

With increasing sample size the BF distributions under $\mathcal{H}_1$ and $\mathcal{H}_0$ diverge (see Figure 3B), making it more likely to obtain compelling evidence for either hypothesis. Consequently, the probability of misleading evidence and the probability of inconclusive evidence is reduced. At $n = 100$ and evidential thresholds of 6 and 1/6 the rate of false negative evidence drops from 0.3% to virtually 0%, and the rate of false positive evidence drops from 0.9% to 0.6%. The probability to detect an existing effect of $\delta = 0.5$ increases from 21.1% to 84.0%, and the probability to find evidence in favor of a true $\mathcal{H}_0$ increases from 13.7% to 53.4%.

Sample size determination. For sample size determination, simulated sample sizes can be adjusted until the com-

[5] The R code is also available on the OSF (https://osf.io/qny5x/).

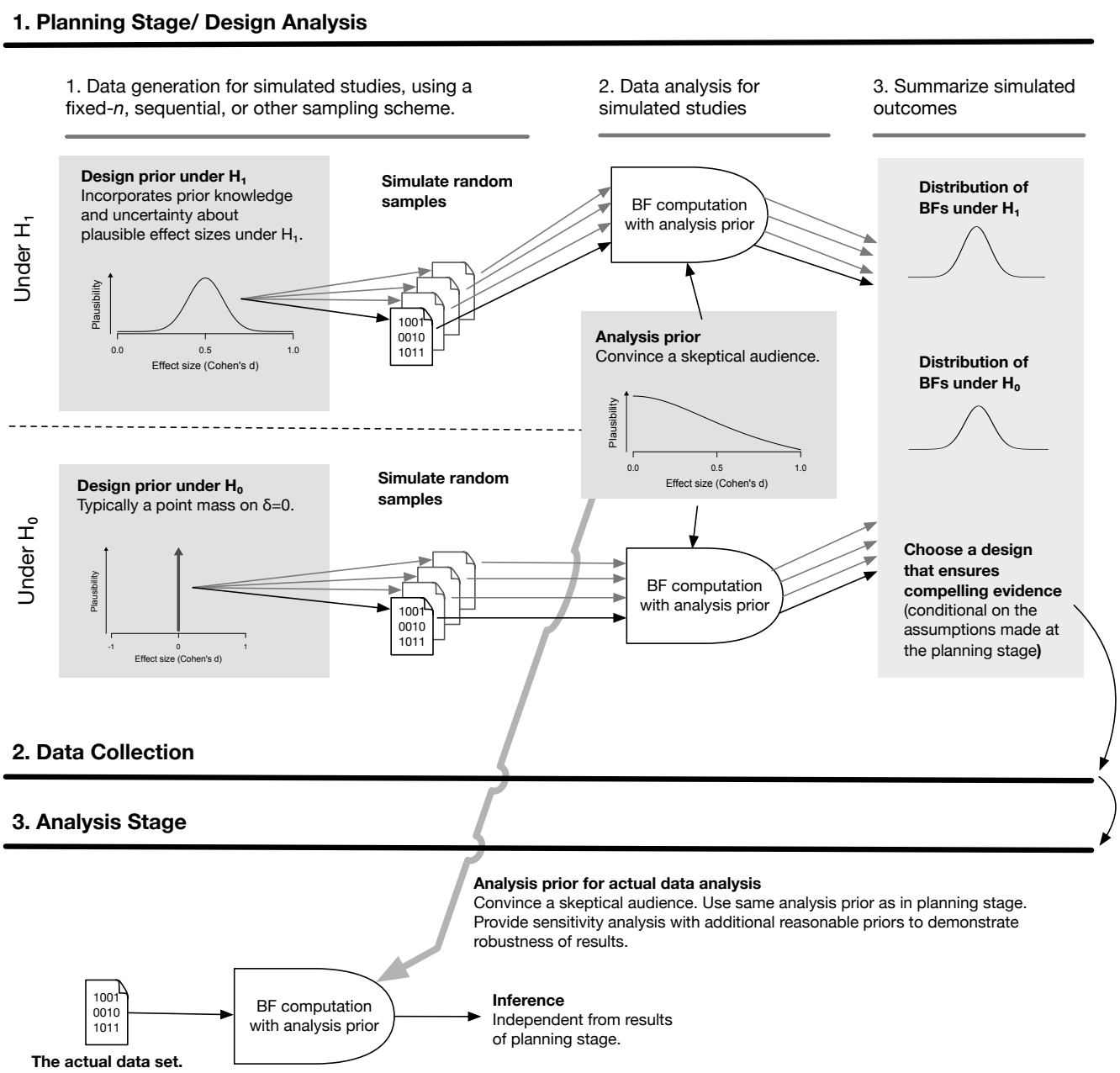

Figure 2. Overview of the planning and analysis stages in a Monte Carlo Bayes factor design analysis. Figure available at https://osf.io/qny5x, under a CC-BY4.0 license.

puted probability of achieving a research goal under $\mathcal{H}_1$ is close to the desired level. In our example, the necessary sample size of achieving a $BF_{10} > 6$ under $\mathcal{H}_1$ with a probability of 95% would be $n = 146$. Such a fixed-n Bayes factor design with $n = 146$ implies a false negative rate of virtually 0%, and, under $\mathcal{H}_0$, a false positive rate of 0.4% and a probability of 61.5% to correctly support $\mathcal{H}_0$.

In a pre-data design perspective the focus is on the frequentist properties of BFs. We should mention that this can be complemented by investigating the Bayesian properties of BFs. From that perspective, one can look at the probability of a hypothesis being true given a certain BF (Rouder, 2014). When $\mathcal{H}_1$ and $\mathcal{H}_0$ have equal prior probability, and when the analysis prior equals the design prior, then a single study with a BF_{10} of, say, 6 has 6:1 odds of stemming from $\mathcal{H}_1$.

The goal of obtaining strong evidence can be achieved by planning a sample size that ensures a strong enough BF with sufficient probability. There is, however, an easier way

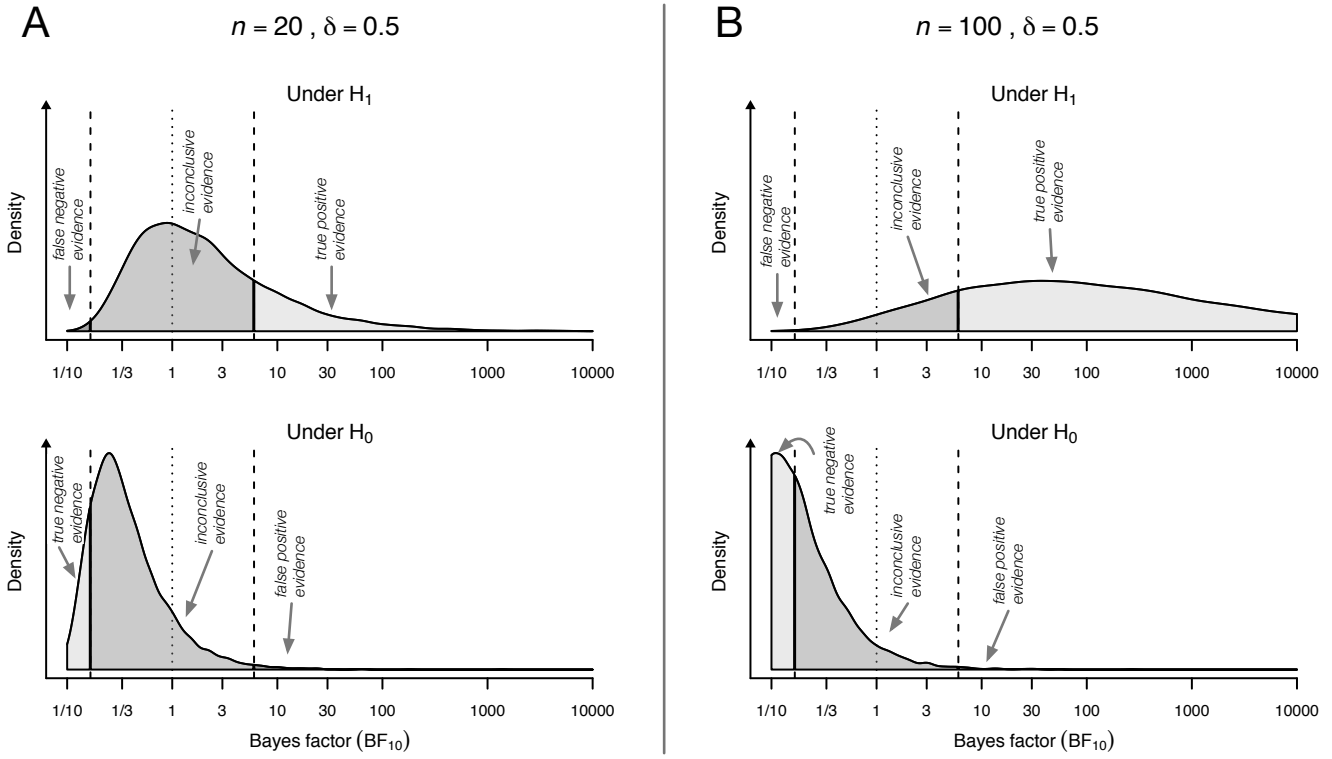

Figure 3. Distributions of BF_{10} for a fixed-*n* design with a true effect size of $\delta = 0.5$ under $\mathcal{H}_1$ and a fixed *n* of 20 (left column), resp. 100 (right column). Distributions were categorized at BF thresholds of 1/6 and 6. Figure available at https://osf.io/qny5x, under a CC-BY4.0 license.

that *guarantees* compelling evidence: Sample sequentially and compute the BF until the desired level of evidence is achieved. This design will be explained in the next section.

Open-ended Sequential Bayes Factor Design: *SBF*

In the planning phase of an experiment, it is often difficult to decide on an expected or minimally interesting effect size. If the planned effect size is smaller than the true effect size, the fixed *n* will be inefficient. More often, presumably, the effect size is overestimated in the planning stage, leading to a smaller actual probability to detect a true effect.

A proposed solution that is less dependent on the true effect size is the Sequential Bayes Factor (SBF) design (Schönbrodt et al., 2015). In this design, the sample size is increased until the desired level of evidence for $\mathcal{H}_1$ or $\mathcal{H}_0$ has been reached (see also Berger, Brown, & Wolpert, 1994, Dienes, 2008, Kass & Raftery, 1995, Lindley, 1956, Wald, 1945). This principle of "accumulation of evidence" is also central to optimal models for human perceptual decision making (e.g., random walk models, diffusion models; e.g., Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006, Forstmann, Ratcliff, & Wagenmakers, 2016). This accumulation principle allows a flexible adaption of the sample size based on the actual empirical evidence.

In the planning phase of a SBF design, researchers define an a priori threshold that represents the desired grade of evidence, for example a BF_{10} of 6 for $\mathcal{H}_1$ and the reciprocal value of 1/6 for $\mathcal{H}_0$. Furthermore, an analysis prior for the effect sizes under $\mathcal{H}_1$ is defined in order to compute the BF. Finally, the researcher may determine a minimum number of participants to be collected regardless, before the optional stopping phase of the experiment (e.g., $n_{min} = 20$ per group).

After a sample of n_{min} participants has been collected, a BF is computed. If this BF does not exceed the $\mathcal{H}_1$ threshold or the $\mathcal{H}_0$ threshold, the sample size is increased as often as desired and a new BF computed at each stage (even after each participant). As soon as one of the thresholds is reached or exceeded, sampling can be stopped. One prominent advantage of sequential designs is that sample sizes are in most cases smaller than those from fixed-*n* designs with the same error rates.[6] For example, in typical scenarios the SBF design for comparing two group means yielded about 50% smaller

[6]For a procedure related to the SBF, the sequential probability ratio test (SPRT; Wald, 1945), it has been proven that this test of two simple (point) hypotheses is an optimal test. That means that of all tests with the same error rates it requires the fewest observations on average (Wald & Wolfowitz, 1948), with sample sizes that are typically 50% lower than the best competing fixed-*n* design.

samples on average compared to the optimal NHST fixed-n design with the same error rates (Schönbrodt et al., 2015).

With regard to design analysis in a SBF design, one can ask: (a) What is the probability of obtaining misleading evidence by stopping at the wrong threshold? (b) What is the expected sample size until an evidential threshold is reached?

In the example for the SBF design, we use a design prior for the a priori effect size estimate: $d \sim \mathcal{N}(0.5, \sigma = 0.1)$ (see Figure 1). In our hypothetical scenario this design prior is inspired by relevant substantive knowledge or results from the published literature. Again, Monte Carlo simulations were used to examine the operational characteristics of this design:

1. Define a population that reflects the expected effect size under $\mathcal{H}_1$ and, if prior information is available, other properties of the real data. In the example given below, we used two populations with normal distributions and a standardized mean difference that has been drawn from a normal distribution $\mathcal{N}(0.5, \sigma = 0.1)$ at each iteration.

2. Draw a random sample of size n_{min} from the populations.

3. Compute the BF for that simulated data set, using the analysis prior that will also be used in the actual data analysis (in our example: a Cauchy prior with scale parameter = $\sqrt{2}/2$). If the BF exceeds the $\mathcal{H}_1$ or the $\mathcal{H}_0$ threshold (in our example: > 6 or $< 1/6$), stop sampling, and save the final BF and the current sample size. If the BF does not exceed a threshold yet, increase sample size (in our example: by 1 in each group). Repeat step 3 until one of both thresholds is exceeded.

4. Repeat steps 1 to 3, say, 10,000 times.

5. In order to compute the rate of false-positive evidence and the expected sample size under $\mathcal{H}_0$, the same simulation must be done under the $\mathcal{H}_0$ (i.e., two populations that have no mean difference).

This design can completely eliminate weak evidence, as data collection is continued until evidence is conclusive in either direction. The consistency property ensures that BFs ultimately drift either towards 0 or towards ∞ and every study ends up producing compelling evidence – unless researchers run out of time, money, or participants (Edwards, Lindman, & Savage, 1963). We call this design "open-ended" because there is no fixed termination point defined a priori (in contrast to the SBF design with maximal sample size, which is outlined below). "Open-ended", however, does not imply that data collection can continue forever without hitting a threshold; in contrast, the consistency property of BFs guarantees that the possibility of collecting samples indefinitely is zero.

Figure 4 (top) visualizes the evolution of the BF_{10} in several studies where the true effect size follows the prior distribution displayed in Figure 1. Each grey line in the plot shows how the BF_{10} of a specific study evolves with increasing n. Some studies hit the (correct) $\mathcal{H}_1$ boundary sooner, some later, and the distribution of stopping-ns is visualized as the density on top of the $\mathcal{H}_1$ boundary. Although all trajectories are guaranteed to drift towards and across the correct threshold in the limiting case, some hit the wrong $\mathcal{H}_0$ threshold prematurely. Most misleading evidence happens at early stages of the sequential sampling. Consequently, increasing n_{min} also decreases the rate of misleading evidence (Schönbrodt et al., 2015). Figure 4 (bottom) shows the same evolution of BFs under $\mathcal{H}_0$.

Expected rates of misleading evidence. If one updates the BF after each single participant under this $\mathcal{H}_1$ of $d \sim \mathcal{N}(0.5, \sigma^2 = 0.1^2)$ and evidential thresholds at 6 and 1/6, 97.2% of all studies stop at the correct $\mathcal{H}_1$ threshold (i.e., the true positive rate), 2.8% stop incorrectly at the $\mathcal{H}_0$ threshold (i.e., the false negative rate). Under the $\mathcal{H}_0$, 93.8% terminate at the correct $\mathcal{H}_0$ threshold, and 6.2% at the incorrect $\mathcal{H}_1$ threshold (i.e., the false positive rate).

The algorithm above computes the BF after each single participant. The more often a researcher checks whether the BF has exceeded the thresholds, the higher the probability of misleading evidence, because the chances are increased that the stop is at a random extreme value. In contrast to NHST, however, where the probability of a Type-I error can be pushed towards 100% if enough interim tests are performed (Armitage, McPherson, & Rowe, 1969), the rate of misleading evidence has an upper limit in the SBF design. When the simulations are conducted with interim tests after each single participant, one obtains the upper bound on the rate of misleading evidence. In the current example this leads to a maximal FPE rate of 6.2%. If the BF is computed after every 5 participants, the rate is reduced to 5.2%, after every 10 participants to 4.5%. It should be noted that these changes in FPE rate are, from an inferential Bayesian perspective, irrelevant (Rouder, 2014).

Expected sample size. In the above example, the average sample size at the stopping point (across both threshold hits) under $\mathcal{H}_1$ is $n = 53$, the median sample size is $n = 36$, and 80% of all studies stop with fewer than 74 participants. Under $\mathcal{H}_0$, the sample size is on average 93, median = 46, and 80% quantile = 115. Hence, although the SBF design has no a priori defined upper limit of sample size, the prospective design analysis reveals estimates of the expected sample sizes.

Furthermore, this example highlights the efficiency of the sequential design. A fixed-n Bayes factor design that also aims for evidence with $BF_{10} \geq 6$ (resp. $\leq 1/6$) with the same true positive rate of 97.2% requires $n = 241$ participants (but will have different rates of misleading evidence).

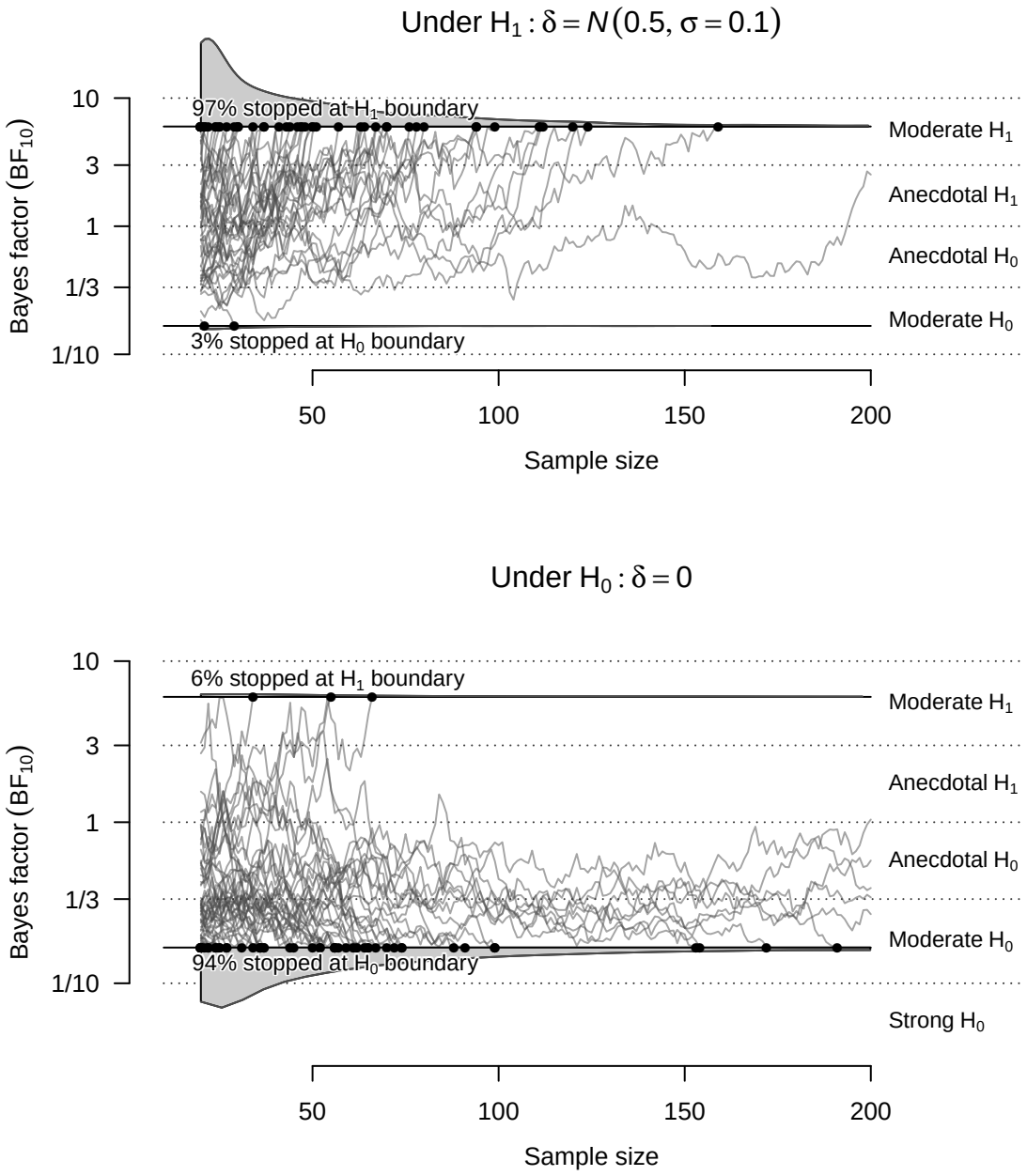

Figure 4. The open-ended Sequential Bayes Factor design. The density of sample sizes at the stopping point and the example trajectories are based on a true effect size of $\delta \sim \mathcal{N}(0.5, \sigma = 0.1)$ under $\mathcal{H}_1$ and evidential thresholds at 6 and 1/6. Figure available at https://osf.io/qny5x, under a CC-BY4.0 license.

Sequential Bayes Factor With Maximal *n*: *SBF+maxN*

The SBF design is attractive because a study is guaranteed to end up with compelling evidence. A practical drawback of the open-ended SBF design, however, is that the BF can meander in the inconclusive region for hundreds or even thousands of participants when effect sizes are very small (Schönbrodt et al., 2015). In practice, researchers do not have unlimited resources, and usually want to set a maximum sample size based on budget, time, or availability of participants.

The *SBF+maxN* design extends the SBF design with such an upper limit on the sample size. Data collection is stopped whenever one of both evidential thresholds has been exceeded, or when the a priori defined maximal sample size has

been reached. When sampling is stopped because $n_{\max}$ has been reached, one can still interpret the final BF. Although it has not reached the threshold for compelling evidence, its direction and strength can still be interpreted.

When planning an *SBF+maxN* design, one can ask: (a) How many studies can be expected to stop because of crossing an evidential threshold, and how many because of reaching $n_{\max}$?, (b) What is the probability of obtaining misleading evidence?, (c) If sampling stopped at $n_{\max}$: How many of these studies have a BF that points into the correct direction? (d) What distribution of sample sizes can be expected?

Again, Monte Carlo simulations can be used to examine the operational characteristics of this design. The computation is equivalent to the SBF design above, with the only exception that step 3 is terminated when the BF exceeds the $\mathcal{H}_1$ or $\mathcal{H}_0$ threshold, or n reaches $n_{\max}$.

To highlight the flexibility and practicality of the *SBF+maxN* design, we consider a hypothetical scenario in which a researcher intends to test as efficiently as possible, has practical limitations on the maximal sample size, and wants to keep the rate of false positive evidence low. To achieve this goal, we introduce some changes to the example from the open-ended *SBF* design above: Asymmetric boundaries, a different minimal sample size, and a maximum sample size.

False positive evidence happens when the $\mathcal{H}_1$ boundary is hit prematurely although $\mathcal{H}_0$ is true. As most misleading evidence happens at early terminations of a sequential design, the FPE rate can be reduced by increasing n_{min} (say, $n_{min} = 40$). Furthermore, the FPE rate can be reduced by a high $\mathcal{H}_1$ threshold (say, $\text{BF}_{10} >= 30$). With an equally strong threshold for $\mathcal{H}_0$ (1/30), however, the expected sample size can easily go into thousands under $\mathcal{H}_0$ (Schönbrodt et al., 2015). To avoid such a protraction, the researcher may set a lenient $\mathcal{H}_0$ threshold of $\text{BF}_{10} < 1/6$. Finally, due to budget restrictions, the maximum affordable sample size is defined as $n_{\max} = 100$. With these settings, the researcher trades in a higher expected rate of false negative evidence (caused by the lenient $\mathcal{H}_0$ threshold), and some probability of weak evidence (when the study is terminated at $n_{\max}$) for a smaller expected sample size, a low rate of false positive evidence and the certainty that the sample size does not exceed $n_{\max}$.

To summarize, in this final example we set evidential thresholds for BF_{10} at 30 and 1/6, $n_{min} = 40$, and $n_{\max} = 100$. The uncertainty about the effect size under $\mathcal{H}_1$ is expressed as $\delta \sim \mathcal{N}(0.5, \sigma = 0.1)$. Figure 5 visualizes the trajectories and stopping point distributions under $\mathcal{H}_1$ (results under $\mathcal{H}_0$ not shown). The upper and lower densities show the distribution of n for all studies that hit a threshold. The distribution on the right shows the distribution of BF_{10} for all studies that stopped at $n_{\max}$.

Expected stopping threshold ($\mathbf{H_1}$, $\mathbf{H_0}$, or $\mathbf{n_{max}}$) and expected rates of misleading evidence. Under $\mathcal{H}_1$ of this example, 70.6% of all studies hit the correct $\mathcal{H}_1$ threshold (i.e., the true positive rate), 1.6% hit the wrong $\mathcal{H}_0$ threshold (i.e, the false negative rate). The remaining 27.8% of studies stopped at $n_{\max}$ and remained inconclusive with respect to the a priori set thresholds.

One goal in the example was a low FPE rate. Under $\mathcal{H}_0$ (not displayed), 70.9% of all studies hit the correct $\mathcal{H}_0$ threshold and 0.6% hit the wrong $\mathcal{H}_1$ threshold (i.e., the false positive rate). The remaining 28.5% of studies stopped at $n_{\max}$ and remained inconclusive with respect to the a priori set thresholds.

Again, these are the maximum rates of misleading evidence, when a test after each participant is computed. More realistic sequential tests, such as testing after every 10 participants, will lower these rates.

Distribution of evidence at $\mathbf{n_{max}}$. The BF of studies that did not reach the a priori threshold for compelling evidence can still be interpreted. In the current example, we categorize the inconclusive studies into results that show at least moderate evidence for either hypothesis ($\text{BF} < 1/3$ or $\text{BF} > 3$) or are completely inconclusive ($1/3 < \text{BF} < 3$). Of course any other threshold can be used to categorize the non-compelling studies; in general a BF of 3 provides only weak evidence for a hypothesis and implies, from a design perspective, a high rate of misleading evidence (Schönbrodt et al., 2015).

In the current example, under $\mathcal{H}_1$, 15.5% of all studies terminated at $n_{\max}$ with a $\text{BF}_{10} > 3$, meaning that these studies correctly indicated at least moderate evidence for $\mathcal{H}_1$. 11.6% of studies remained inconclusive ($1/3 < \text{BF}_{10} < 3$), and 0.7% pointed towards the wrong hypothesis ($\text{BF}_{10} < 1/3$). Under $\mathcal{H}_0$, 1.1% incorrectly pointed towards $\mathcal{H}_1$, 10.8% towards $\mathcal{H}_0$, and 16.6% remained inconclusive.

Expected sample size. The average expected sample size under $\mathcal{H}_1$ (combined across all studies, regardless of the stopping condition) is $n = 69$, with a median of 65. The average expected sample size under $\mathcal{H}_0$ is $n = 66$, with a median of 56. Hence, the average expected sample size is under both hypotheses considerably lower than $n_{\max}$, which has been defined at $n = 100$.

Discussion

We explored the concept of a Bayes Factor Design Analysis, and how it can help to plan a study for compelling evidence. Pre-data design analyses allow researchers to plan a study in a way that strong inference is likely. As in frequentist power analysis, one has to find a trade-off between the rates of misleading evidence, the desired probability of achieving compelling evidence, and practical limits concerning sample size. Additionally, in order to compute the expected outcomes of future studies, one has to make explicit one's assumption for several key parameters, such as the expected effect size under $\mathcal{H}_1$. Any pre-data analysis is conditional on these assumptions, and the validity of the results

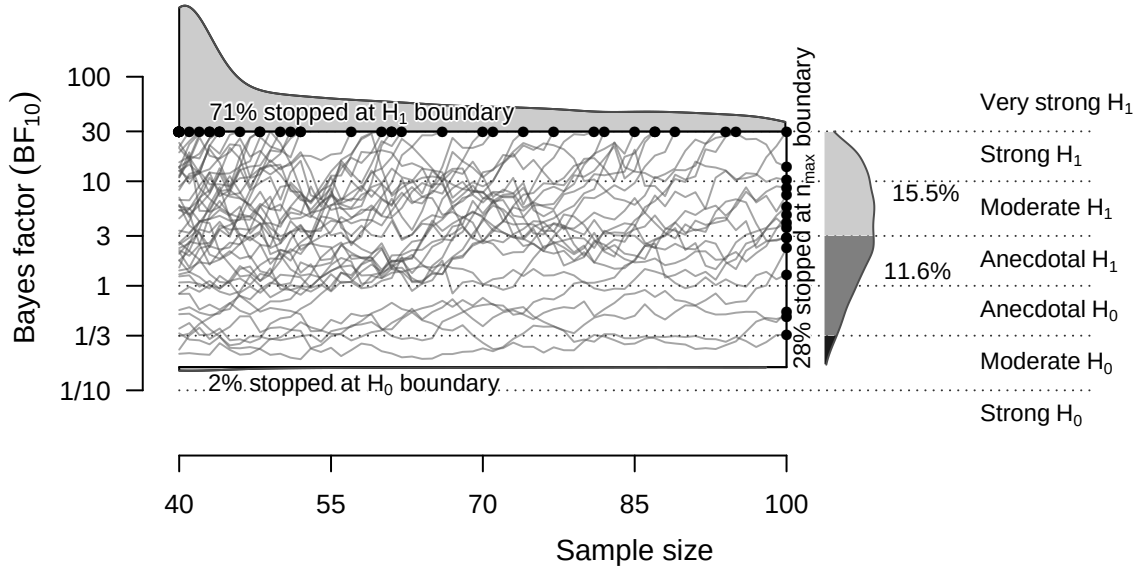

Figure 5. The Sequential Bayes Factor With Maximal n design under $\mathcal{H}_1$ (results under $\mathcal{H}_0$ not shown). The densities and example trajectories are based on a true effect size of $\delta \sim \mathcal{N}(0.5, \sigma = 0.1)$, evidential thresholds at 30 and 1/6, and $n_{\max} = 100$ in each group. Figure available at https://osf.io/qny5x, under a CC-BY4.0 license.

depends on the validity of the assumptions. If reality does not follow the assumptions, the actual operational characteristics of a design will differ from the results of the design analysis. For example, if the actual effect size is smaller than anticipated, a chosen design has actually higher FNE rates and, in the sequential case, larger expected sample sizes until a threshold is reached.

In contrast to p-values, the interpretation of Bayes factors does not depend on stopping rules (Rouder, 2014). This property allows researchers to use flexible research designs without the requirement of special and ad-hoc corrections. For example, the proposed *SBF+maxN* design stops abruptly at $n_{\max}$. An alternative procedure is one where the evidential thresholds gradually move closer together as n increases. This implies that a lower grade of evidence is accepted when sampling was not already stopped at a strong evidential threshold, and puts a practical (but not fixed) upper limit on sample size (for an application in response time modeling see Boehm, Hawkins, Brown, van Rijn, & Wagenmakers, 2015). The properties of this special design (or of any sequential or non-sequential BF design) can be evaluated using the same simulation approach outlined in this paper. This further underscores the flexibility and the generality of the sequential Bayesian procedure.

From the Planning Stage to the Analysis Stage

This paper covered the planning stage, before data are collected. After a design has been chosen, based on a careful evaluation of its operational characteristics, the actual study is carried out (see also Figure 2). A design analysis only relates to the actual inference if the same analysis prior is used in the planning stage and in the analysis stage. Additionally, the BF computation in the analysis stage should contain a sensitivity analysis, which shows whether the inference is robust against reasonable variations in the analysis prior.

It is important to note that, in contrast to NHST, the inference drawn from the actual data set is entirely independent from the planning stage (Berger & Wolpert, 1988; Dienes, 2011; Wagenmakers et al., 2014). All inferential information is contained in the actual data set, the analysis prior, and the likelihood function. Hypothetical studies from the planning stage (that have not been done) cannot add anything. From that perspective, it would be perfectly fine to use a different analysis prior in the actual analysis than in the design analysis. This would not invalidate the inference (as long as the chosen analysis prior is defensible); it just would disconnect the pre-data design analysis, which from a post-data perspective is irrelevant anyway, from the actual analysis.

Unbiasedness of Effect Size Estimates

Concerning the sequential procedures described here, some authors have raised concerns that these procedures result in biased effect size estimates (e.g., Bassler et al., 2010, Kruschke, 2014). We believe these concerns are overstated, for at least two reasons.

First, it is true that studies that terminate early at the $\mathcal{H}_1$ boundary will, on average, overestimate the true effect. This conditional bias, however, is balanced by late terminations, which will, on average, underestimate the true effect. Early terminations have a smaller sample size than late terminations, and consequently receive less weight in a meta-analysis. When all studies (i.e., early and late terminations) are considered together, the bias is negligible (Berry, Bradley, & Connor, 2010; Fan, DeMets, & Lan, 2004; Goodman, 2007; Schönbrodt et al., 2015). Hence, across multiple studies the sequential procedure is approximately unbiased.

Second, the conditional bias of early terminations is conceptually equivalent to the bias that results when only significant studies are reported and non-significant studies disappear into the file drawer (Goodman, 2007). In all experimental designs –whether sequential, non-sequential, frequentist, or Bayesian– the average effect size inevitably increases when one selectively averages studies that show a larger-than-average effect size. Selective publishing is a concern across the board, and an unbiased research synthesis requires that one considers significant and non-significant results, as well as early and late terminations.

Although sequential designs have negligible unconditional bias, it may nevertheless be desirable to provide a principled "correction" for the conditional bias at early terminations, in particular when the effect size of a single study is evaluated. For this purpose, Goodman (2007) outlines a Bayesian approach that uses prior expectations about plausible effect sizes (see also Pocock & Hughes, 1989). This approach shrinks extreme estimates from early terminations towards more plausible regions. Smaller sample sizes are naturally more sensitive to prior-induced shrinkage, and hence the proposed correction fits the fact that most extreme deviations from the true value are found in very early terminations that have a small sample size (Schönbrodt et al., 2015).

Practical Considerations

Many granting agencies require a priori computations for the determination of sample size. This ensures that proposers explicitly consider the expected or minimally relevant effect size. Such calculations are necessary to pinpoint the amount of requested money to pay participants.

The *SBF+maxN* design seems especially suitable for a scenario where researchers want to take advantage of the high efficiency of a sequential design but still have to define a fixed (maximum) sample size in a proposal. For this purpose, one could compute a first design analysis based on an open-ended SBF design to determine a reasonable $n_{\max}$. If, for example, the 80% quantile of the stopping-*n* distribution is used as n_{max} in a *SBF+maxN* design, one can expect to hit a boundary before $n_{\max}$ is reached in 80% of all studies. Although there is a risk of 20% that a study does not reach compelling evidence within the funding limit, this outcome is not a "failure" as the direction and the size of the final BF can still be interpreted. In a second design analysis one should consider the characteristics of that *SBF+maxN* design and evaluate whether the rates of misleading evidence are acceptable.

This approach enables researchers to define an informed upper limit for sample size, which allows them to apply for a predefined amount of money. Still, one can save resources if the evidence is strong enough for an earlier stop, and in almost all cases the study will be more efficient than a fixed-*n* NHST design with comparable error rates (Schönbrodt et al., 2015).

Conclusion

In the planning phase of a study it is essential to carry out a design analysis in order to formalize one's expectations and facilitate the design of informative experiments. A large body of literature is available on planning frequentist designs, but little practical advice exists for research designs that employ Bayes factors as a measure of evidence. In this contribution we elaborate on three BF designs –a fixed-*n* design, an open-ended Sequential Bayes Factor (SBF) design, and an SBF design with maximal sample size– and demonstrate how the properties of each design can be evaluated using Monte Carlo simulations. Based on the analyses of the operational characteristics of a design, the specific settings of the research design can be balanced in a way that compelling evidence is a likely outcome of the to-be-conducted study, misleading evidence is an unlikely outcome, and sample sizes are within practical limits.

References

Armitage, P., McPherson, C. K., & Rowe, B. C. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society. Series A (General)*, *132*(2), 235–244.

Bacchetti, P., Wolf, L. E., Segal, M. R., & McCulloch, C. E. (2005). Ethics and sample size. *American Journal of Epidemiology*, *161*(2), 105–110. doi:10 . 1093 / aje / kwi014

Bassler, D., Briel, M., Montori, V. M., Lane, M., Glasziou, P., Zhou, Q., … Guyatt, G. H. (2010). Stopping randomized trials early for benefit and estimation of treatment effects: Systematic review and meta-regression analysis. *Journal of the American Medical Association*, *303*(12), 1180–1187.

Bayarri, M. J., Benjamin, D. J., Berger, J. O., & Sellke, T. M. (2016). Rejection odds and rejection ratios: A proposal for statistical practice in testing hypotheses. *Journal of Mathematical Psychology*, *72*, 90–103. doi:10.1016/j.jmp.2015.12.007

Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis* (2nd edition). New York: Springer.

Berger, J. O. (2006). Bayes factors. In S. Kotz, N. Balakrishnan, C. Read, B. Vidakovic, & N. L. Johnson (Eds.), *Encyclopedia of statistical sciences, vol. 1 (2nd ed.)* (pp. 378–386). Hoboken, NJ: Wiley.

Berger, J. O. & Wolpert, R. L. (1988). *The likelihood principle (2nd ed.)* Hayward, CA: Institute of Mathematical Statistics.

Berger, J. O., Brown, L. D., & Wolpert, R. L. (1994). A unified conditional frequentist and Bayesian test for fixed and sequential simple hypothesis testing. *The Annals of Statistics*, *22*(4), 1787–1807. doi:10.1214/aos/1176325757

Berry, S. M., Bradley, P. C., & Connor, J. (2010). Bias and trials stopped early for benefit. *JAMA*, (304), 156–159. doi:doi:10.1001/jama.2010.930

Blume, J. D. (2002). Likelihood methods for measuring statistical evidence. *Statistics in Medicine*, *21*(17), 2563–2599. doi:10.1002/sim.1216

Blume, J. D. (2008). How often likelihood ratios are misleading in sequential trials. *Communications in Statistics: Theory & Methods*, *37*(8), 1193–1206. doi:10.1080/03610920701713336

Boehm, U., Hawkins, G. E., Brown, S., van Rijn, H., & Wagenmakers, E.-J. (2015). Of monkeys and men: Impatience in perceptual decision-making. *Psychonomic Bulletin & Review*, *23*(3), 738–749. doi:10.3758/s13423-015-0958-5

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The Physics of Optimal Decision Making: A Formal Analysis of Models of Performance in Two–alternative Forced Choice Tasks. *Psychological Review*, *113*, 700–765.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376.

Cavagnaro, D. R., Myung, J. I., Pitt, M. A., & Kujala, J. V. (2009). Adaptive design optimization: A mutual information-based approach to model discrimination in cognitive science. *Neural Computation*, *22*(4), 887–905. doi:10.1162/neco.2009.02-09-959

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New Jersey, US: Lawrence Erlbaum Associates.

De Santis, F. (2004). Statistical evidence and sample size determination for Bayesian hypothesis testing. *Journal of Statistical Planning and Inference*, *124*, 121–144.

Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. New York: Palgrave Macmillan.

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, *6*(3), 274–290. doi:10.1177/1745691611406920

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology: Quantitative Psychology and Measurement*, *5*, 781. doi:10.3389/fpsyg.2014.00781

Dienes, Z. (2016). How Bayes factors change scientific practice. *Journal of Mathematical Psychology*, *72*, 78–89. doi:10.1016/j.jmp.2015.10.003

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., … Johannesson, M. (2015). Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences*, *112*(50), 15343–15347. doi:10.1073/pnas.1516179112

Eaton, M. L., Muirhead, R. J., & Soaita, A. I. (2013). On the limiting behavior of the "probability of claiming superiority" in a Bayesian context. *Bayesian Analysis*, *8*(1), 221–232.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*(3), 193–242. doi:10.1037/h0044139

Emanuel, E. J., Wendler, D., & Grady, C. (2000). What makes clinical research ethical? *JAMA*, *283*(20), 2701–2711.

Fan, X., DeMets, D. L., & Lan, K. K. G. (2004). Conditional bias of point estimates following a group sequential test. *Journal of Biopharmaceutical Statistics*, *14*(2), 505–530. doi:10.1081/BIP-120037195

Forstmann, B. U., Ratcliff, R., & Wagenmakers, E.-J. (2016). Sequential Sampling Models in Cognitive Neuroscience: Advantages, Applications, and Extensions. *Annual Review of Psychology*, *67*, 641–666.

Garthwaite, P. H., Kadane, J. B., & O'Hagan, A. (2005). Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, *100*(470), 680–701. doi:10.1198/016214505000000105

Gelman, A. & Carlin, J. (2014). Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science*, *9*(6), 641–651. doi:10.1177/1745691614551642

Gelman, A. & Tuerlinckx, F. (2000). Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics*, *15*(3), 373–390.

Good, I. J. (1979). Studies in the history of probability and statistics. XXXVII A. M. Turing's statistical work in World War II. *Biometrika*, *66*(2), 393–396. doi:10.1093/biomet/66.2.393

Goodman, S. N. (2007). Stopping at nothing? Some dilemmas of data monitoring in clinical trials. *Annals of Internal Medicine*, *146*(12), 882–887.

Halpern, S. D., Karlawish, J. H. T., & Berlin, J. A. (2002). The continuing unethical conduct of underpowered clinical trials. *JAMA*, *288*(3), 358–362.

Hoijtink, H., Klugkist, I., & Boelen, P. (2008). *Bayesian Evaluation of Informative Hypotheses*. New York: Springer.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Med*, *2*(8), e124. doi:10.1371/journal.pmed.0020124

JASP Team, T. (2016). JASP (Version 0.7.5.6)[Computer software].

Jeffreys, H. (1961). *The theory of probability*. Oxford University Press.

Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

Kruschke, J. K. (2010). Bayesian data analysis. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(5), 658–676. doi:10.1002/wcs.72

Kruschke, J. K. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd edition). Boston: Academic Press.

Lakens, D. & Evers, E. R. K. (2014). Sailing from the seas of chaos into the corridor of stability: Practical recommendations to increase the informational value of studies. *Perspectives on Psychological Science*, *9*(3), 278–292. doi:10.1177/1745691614528520

Lee, M. D. & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.

Lewis, S. M. & Raftery, A. E. (1997). Estimating Bayes Factors via posterior simulation with the Laplace–Metropolis estimator. *Journal of the American Statistical Association*, *92*, 648–655.

Lindley, D. V. (1956). On a Measure of the Information Provided by an Experiment. *The Annals of Mathematical Statistics*, *27*, 986–1005.

Lindley, D. V. (1997). The choice of sample size. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *46*(2), 129–138.

Ly, A., Verhagen, J., & Wagenmakers, E.-J. (2016). Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology*. Bayes Factors for Testing Hypotheses in Psychological Research: Practical Relevance and New Developments, *72*, 19–32. doi:10.1016/j.jmp.2015.06.004

Morey, R. D. & Rouder, J. N. (2015). *BayesFactor: Computation of Bayes factors for common designs*.

Morey, R. D., Romeijn, J.-W., & Rouder, J. N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*. Bayes Factors for Testing Hypotheses in Psychological Research: Practical Relevance and New Developments, *72*, 6–18. doi:10.1016/j.jmp.2015.11.001

Morris, D. E., Oakley, J. E., & Crowe, J. A. (2014). A web-based tool for eliciting probability distributions from experts. *Environmental Modelling & Software*, *52*, 1–4. doi:10.1016/j.envsoft.2013.10.010

Muirhead, R. J. & Şoaita, A. I. (2013). On an approach to Bayesian sample sizing in clinical trials. In G. Jones & X. Shen (Eds.), *Advances in Modern Statistical Theory and Applications: A Festschrift in honor of Morris L. Eaton* (pp. 126–137). Beachwood, Ohio: Institute of Mathematical Statistics.

Mulder, J. & Wagenmakers, E.-J. (2016). Editors' introduction to the special issue "Bayes factors for testing hypotheses in psychological research: Practical relevance and new developments". *Journal of Mathematical Psychology*. Bayes Factors for Testing Hypotheses in Psychological Research: Practical Relevance and New Developments, *72*, 1–5. doi:10.1016/j.jmp.2016.01.002

O'Hagan, A. & Forster, J. (2004). *Kendall's Advanced Theory of Statistics Vol. 2B: Bayesian Inference (2nd ed.)* London: Arnold.

O'Hagan, A. & Stevens, J. W. (2001 May-Jun). Bayesian assessment of sample size for clinical trials of cost-effectiveness. *Medical Decision Making: An International Journal of the Society for Medical Decision Making*, *21*(3), 219–230.

O'Hagan, A., Stevens, J. W., & Campbell, M. J. (2005). Assurance in clinical trial design. *Pharmaceutical Statistics*, *4*(3), 187–201. doi:10.1002/pst.175

Platt, J. R. (1964). Strong inference. *Science*, *146*(3642), 347–353. doi:10.1126/science.146.3642.347

Pocock, S. J. & Hughes, M. D. (1989). Practical problems in interim analyses, with particular regard to estimation. *Controlled Clinical Trials*, *10*(4), 209–221.

R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria.

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*(2), 301–308. doi:10.3758/s13423-014-0595-4

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA de-

signs. *Journal of Mathematical Psychology*, *56*(5), 356–374. doi:10.1016/j.jmp.2012.08.001

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237.

Royall, R. M. (2000). On the probability of observing misleading statistical evidence. *Journal of the American Statistical Association*, *95*(451), 760–768. doi:10.2307/2669456

Schönbrodt, F. D. (2016). BFDA: Bayes factor design analysis package for R. https://github.com/nicebread/BFDA.

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2015). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*. doi:10.1037/met0000061

Spiegelhalter, D. J., Abrams, K. R., & Myles, J. P. (2004). *Bayesian approaches to clinical trials and health-care evaluation*. John Wiley & Sons.

Taroni, F., Bozza, S., Biedermann, A., Garbolino, P., & Aitken, C. (2010). *Data analysis in forensic science: A Bayesian decision perspective*. Chichester: John Wiley & Sons.

van Erven, T., Grünwald, P., & de Rooij, S. (2012). Catching up faster by switching sooner: A predictive approach to adaptive estimation with an application to the AIC–BIC dilemma. *Journal of the Royal Statistical Society B*, *74*, 361–417.

Wagenmakers, E.-J., Grünwald, P., & Steyvers, M. (2006). Accumulative prediction error and the selection of time series models. *Journal of Mathematical Psychology*, *50*, 149–166.

Wagenmakers, E.-J., Morey, R. D., & Lee, M. D. (2016). Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science*, *25*(3), 169–176. doi:10.1177/0963721416643289

Wagenmakers, E.-J., Verhagen, J., Ly, A., Bakker, M., Lee, M. D., Matzke, D., ... Morey, R. D. (2014). A power fallacy. *Behavior Research Methods*, *47*(4), 913–917. doi:10.3758/s13428-014-0517-4

Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, *16*(2), 117–186.

Wald, A. & Wolfowitz, J. (1948). Optimum character of the sequential probability ratio test. *The Annals of Mathematical Statistics*, *19*(3), 326–339. doi:10.1214/aoms/1177730197

Walley, R. J., Smith, C. L., Gale, J. D., & Woodward, P. (2015). Advantages of a wholly Bayesian approach to assessing efficacy in early drug development: a case study. *Pharmaceutical Statistics*, *14*(3), 205–215. doi:10.1002/pst.1675

Weiss, R. (1997). Bayesian sample size calculations for hypothesis testing. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *46*(2), 185–191.