LUNDS
UNIVERSITET

DEPARTMENT of PSYCHOLOGY

Sequential analyses in psychological research using Bayesian statistics

Pierre Klintefors

Master's Thesis (15 hp)
Spring 2019

Supervisor: Geoffrey Patching

Abstract

It is important in psychological research to use well planned methods that are as time and resource efficient as possible, without jeopardizing the reliability and validity of psychological science. The present paper aims to test how sequential analyses could be implemented in psychological research using Bayesian statistics. With sequential analyses it is possible to stop an experiment or study in the data collection stage for success or futility. To avoid offset estimation and false alarms, a mixture of model testing with Bayes Factor and Bayesian parameter estimation were used as stopping rules. After several runs of Monte Carlo simulations, it appears as a Bayes' Factor (BF) boundary of 6 together with 95% Highest density interval (HDI) width under a *SD**0.60 served as suitable stopping rules under conditions of simulations. However, the generalizability is limited by the simulations settings and the stopping rules are recommended to be implemented on data from real conducted experiments.

Sequential analyses in psychological research using Bayesian statistics

Psychological scientists aim to conduct studies and experiments with good reliability and validity in order to draw credible conclusions that address their research questions. However, researchers do not have unlimited time and resources, instead they try to conduct the best studies or experiments possible with the means available to them. Therefore, it is important to use well planned methods that are as time and resource efficient as possible, without jeopardizing the reliability and validity of psychological science. One potential way of increasing the efficiency of psychological research is to test the data after every participant with the aim of obtaining compelling evidence with the smallest sample size possible (Lakens, 2014; Kruschke, 2012; Schönbrod & Wagenmakers, 2018). The objective of the present paper is to explore ways to implement sequential analyses in psychological science in order to improve the efficiency of data collection and hypothesis testing. The operating characteristics of interim Bayesian data analysis are investigated as an alternative to frequentist Null Hypothesis Significance Testing (NHST) for sequential analyses of psychological data.

By way of sequential Bayesian testing it is possible to conduct an experiment without predetermination of a fixed sample size, but instead stop data collection based on the estimated effect size or precision of the parameter estimates, a procedure known as optional stopping (Armitage, McPherson, & Rowe, 1969). Sequential testing and optional stopping are known to reduce both the cost and time required for psychological studies (Lai, 1973; Lakens, 2014; Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, 2015). Optional stopping is thereby an attractive selling point in the recruitment of financial resources. It can be used to ensure potential financiers that precautionary actions are in place to reduce the risk of wasting resources. Beyond practical or economic benefits of using sequential analysis for optional stopping there are also ethical arguments by which to recommend this approach. Many people would agree that it is unethical to expose more participants to experimental conditions than strictly necessary. This is especially true if the experimental manipulation poses some risk to participants, such as Transcranial Magnetic Stimulation. The present paper sets out to examine sequential analyses and assess different decision rules for optional stopping by which to potentially increase the efficiency of psychological science. The stopping rules assessed in the present paper concern Bayes' Factor (BF) analyses with specified thresholds along with Bayesian parameter estimation. The framework of using sequential BF analysis is a lively area of discussion in the psychological

literature (Morey & Rouder, 2011; Rouder, Morey, Verhagen, Province, & Wagenmakers, 2016; Wagenmakers, Lee, Lodewyckx, & Iverson, 2008; Schönbrodt, Wagenmakers, 2018) but, beyond brief discussion by Kruschke (2011, 2015), sequential Bayesian parameter estimation has received far less attention.

Hypothesis testing in psychology

The most common statistical approach in psychology is a mixture between Fisher's (1925) approach to Null Hypothesis testing and the Neyman-Pearson (1933) approach of stating an alternative hypothesis with a decision bound to reject the Null hypothesis based on a predefined *alpha* level. In psychology the alpha level is usually set to 5% to keep the Type I error rate within acceptable limits. Type I error refers to the statistical decision error of stating that an effect exists when it does not. According to Neyman-Pearson approach, it is necessary to specify a Null hypothesis, which is a prediction of no effect, and an alternative hypothesis, which is the prediction of an effect of interest. The idea is to collect sufficient data to determine, with an adequate level of certainty, whether there is an effect by looking at how likely a test statistic is given that the Null hypothesis is true. Most usually, this is done by drawing a sample from the population of interest to estimate the parameter(s) of interest in that population. The larger the sample size the more likely the sample will be representative of the population from which it is drawn and increase the likelihood of finding hypothesized effects if they are true. When planning a study, it is important to assure that it has sufficient 'statistical power'. Statistical power is the probability that a statistical test will show significant results given that the alternative hypothesis is true (Ellis, 2010). Statistical power is based on four quantities: (1) the estimated size of the effect, (2) the estimated variance of the data, (3) the Type I error rate, and (4) the sample size. Most often, the size of the effect and the variance of the data are combined into a standardized effect size (Cohen, 1988). By convention the Type I error rate in psychology is set at 5%, although the error rate can be set lower (i.e., 1% or 0.1%) if more compelling evidence is required. After choosing a specific value for the hypothesized standardized effect size in the population, and the Type I error rate specified, estimation of a-priori statistical power depends only on sample size. A-priori power analysis, therefore, provides an estimation of the required sample size to obtain a statistically significant result given that the effect size exactly specified truly exist in the population from which the sample is drawn. A-priori power analysis serves as a reasonable tool for predetermination of sample size, for so-called fixed-*n* designs. A fixed-*n*

design refers to a study design in which the number of participants to be tested is decided on in advance of the study, optimally on the basis of statistical power analysis.

A well-known problem with a-priori NHST power analysis is that interpretation of *p*-values depends on the precise testing intentions of the researcher (Lakens, 2014); any deviation from the preplanned testing schedule can dramatically increase the Type I error rate or decrease the power of the experiment. Fixed-*n* designs based on a-priori power analysis depend exactly on how close the pre-specified effect size is to the true effect size within the population. If the true effect size in the population is smaller than expected then the planned study may not have sufficient power to reliably detect the true effect in the population (Schönbrodt, Wagenmakers, Zehetleitner, & Perugin, 2017). Alternatively, if the precisely pre-specified effect size for a-priori power analysis is smaller than the true effect size in the population, running more participants than necessary to detect the true population effect size is inevitably a waste of time and resources. A further problem arises with fixed-*n* designs when marginally significant results obtain after the exactly planned number of participants are tested. In this case, it is not statistically acceptable to run more participants due to problems associated with deviating from the pre-planned testing schedule and inflation of the Type I error rate.

Sequential analysis

As an alternative to standard NHST with a fixed-*n* design and with the power of modern desktop computers, data analyses may be conducted continually as the experiment is being run. Sequential analysis, sometimes called sequential hypothesis testing, refers to the practice of conducting interim data analyses while data collection is in process. On the basis of sequential testing there is potentially no need for a fixed-*n* sample size, instead sequential testing introduces the possibility of optional stopping, where the experiment is terminated or continued on the basis of the analysis (Armitage, McPherson, & Rowe, 1969; Lai, 1973; Lakens, 2014). In this regard, a variety of stopping rules can be identified, depending on the goals or resources of the researcher. A stopping rule may, for example, be based on obtaining a minimal effect size of interest along with a reasonably small confidence interval (CI) that are considered suitably precise in relation to the research question of interest; termed, stopping for success. Sequential analyses are frequently used in clinical research where it is of great importance to reduce the amount of unnecessary risk exposure for participants (Freedman, Lowe, & Macaskill, 1984). By sequentially analyzing the data as it is being collected it is also possible to terminate an experiment if the analysis shows

that it is highly unlikely that the desired effect size will be obtained, given the time and resources available; termed, stopping for futility. In this regard, sequential data analysis allows for the possibility of stopping for futility or success depending on the goals or resources of the researcher, providing for a more efficient study design than the traditional fixed-*n* design (Lakens 2014, Schönbrodt, et al., 2017).

Sequential analysis in psychology. Sequential analyses, or 'peeking' at data as it is being collected, is possibly common practice in psychology (John, Loewenstein, & Prelec, 2012; Yu, Sprenger, Thomas, & Dougherty, 2014) but rarely admitted due to problems associated with inflating the NHST Type I error rate (Armitage, McPherson, & Rowe, 1969, Lakens, 2014; Proschan, Lan, & Wittes, 2006; Simmons, Nelson, & Simonsohn, 2011). Given the frequentist NHST framework the Type I error rate dramatically increases with multiple significance testing. Most researchers in psychology are familiar with this problem when multiple significance tests are conducted, and reviewers then demand that the alpha level is adjusted accordingly. If frequentist significance testing is repeated multiple times a *p*-value < .05 will sooner or later be obtained, even if there is no true population difference between the tested effects, leading to an unacceptably high false alarm rate if the alpha level is not adjusted accordingly. Indeed, sequential testing until $p < .05$ obtains is notoriously referred to as "sampling to reach a foregone conclusion" (after Anscombe, 1954). This can be illustrated by way of computer simulation where a very large number of random values (say 2,000,000) are drawn from a normal distribution with a mean of 0 and a standard deviation of 1 to create two similar, hypothetical, populations whereby the Null hypothesis is true. The detailed code for this simulation named can be found on Github-link below[1]. The two independent groups may then be sampled from these hypothetical populations in which the Null hypothesis is predefined as true. The simulation starts by drawing two samples from each hypothetical population to form two groups (so the minimal sample size in each group is 2). If a *p*-value obtained by way of a two sample independent *t* test is less than .05 the simulation is terminated, and the sample size recorded. Alternatively, if the *p*-value is greater than .05 an additional sample from each hypothetical population is added to each group, and the testing procedure is repeated, until $p < .05$ obtains or a maximal sample size is reached. The results of this simulation are shown in Figure 1. When the sample size of each

[1] https://github.com/pierreklintefors/MasterThesis/blob/master/PvalueFA_Optim.R

group reaches n = 2500, nearly 60% of the simulated sequences show a statistically significant difference between the two groups, even though the Null hypothesis is true. In the limit, the false-alarm rate tends towards 100%.

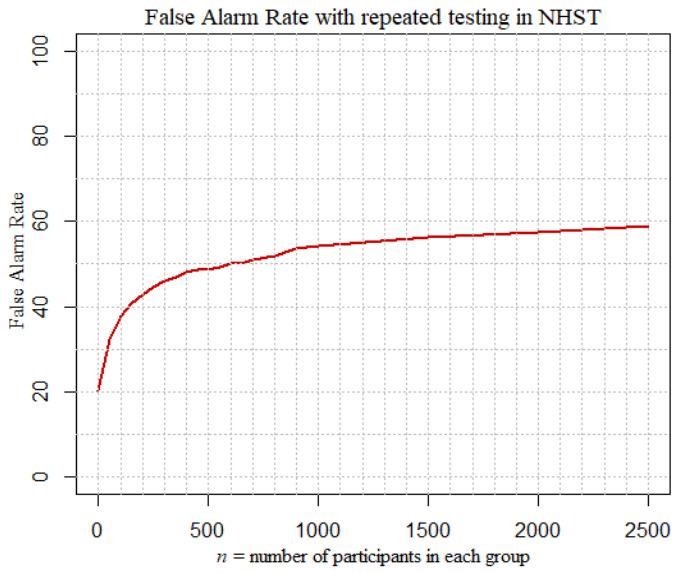

Figure. 1. Proportion of statistically significant results ($p < .05$) obtained by way of sequentially testing two independent groups repeatedly sampled from two hypothetical populations in which the Null hypothesis is true. The Type I error rate is a monotonically increasing function of the sequential testing procedure, tending to 100% in the limit.

The current simulation clearly shows that if sequential analysis is conducted following a frequentist NHST approach, the *alpha*-level must be adjusted accordingly depending on the planned maximal sample size and testing intentions of the researcher (Lakens, 2014). So, it is possible to perform sequential analysis using frequentist NHST procedures, but this is very rarely undertaken due to difficulties associated with implementing correct procedures. When a new medical device needs approval in medical research, a clinical trial is overseen by a committee who meet at regular intervals to review the trial, and ensure that the testing schedule is adhered to strictly as planned (Berry, Carling, Lee, & Müller, 2011). However, in psychology this is usually not financially or practically possible. Psychological studies are typically conducted by lone researchers or small groups of researchers who for practical reasons may not be able to adhere strictly to a planned testing schedule, inadvertently increasing the Type I rate or

decreasing statistical power. A further problem with using the frequentist NHST approach in sequential analysis is that it is not possible to revise the testing plan once started. If marginally significant results obtain, once the maximum pre-planned number of participants is tested, there is nothing more that can be done to obtain compelling data.

Bayesian Statistics

Bayesian statistics can be used as an alternative to frequentist statistics in psychology. To comprehend how Bayesian statistics work it is necessary to understand some basic rules of probability. In frequentist statistics, probability is defined purely in terms of the long run frequency of a random event, such as the probability of an outcome of a role of a dice or flip of a coin. With Bayesian statistics, probability is not restricted to long run frequency, Bayesian statistics allows for defining probabilities for statements or propositions that can be based on long-run frequency or on knowledge about the probability of events (Koch, 2007). Bayesian statistics incorporates probabilities of events based on previous knowledge, such as the probability that it will rain tomorrow. The probability of rain tomorrow can be based on knowledge of the weather patterns in previous days and collected meteorological data of air pressure, temperature, cloudiness, and so forth.

Probability and Bayes' theorem. Probability expresses plausibility and so can be seen as a measure of the plausibility of statements. Bayesian statistics is founded on Bayes' theorem, or Bayes' rule, and work on inverse probability by Pierre Simon de Laplace (1774), where conditional probabilities of events given certain circumstances are computed (Debnath & Basu, 2015; Laplace, 1774). A conditional probability is the probability of one event given another, which in terms of a hypothesis, H, and data, D, is most normally denoted $P(H|D)$, read as the probability of the hypothesis given the data (Koch, 2007). For example, the probability (P) that it will rain tomorrow (H), given that it rained today (D). Not always but most usually, $P(D|H) \neq P(H|D)$, the probability that it will rain tomorrow given that it rained today is not the same as the probability that it rained today given that it will rain tomorrow. Bayes' rule is a mathematical formula that takes us from one conditional probability, $P(D|H)$, to the other, ($P(H|D)$, and is denoted as:

$$P(H|D) = \frac{P(H)\,P(D|H)}{P(D)}$$

Bayes' rule gives us the relationship between two conditional probabilities and makes it possible to get posterior knowledge, $P(H|D)$, based on prior knowledge $P(H)$ as well as the collected data, $P(D|H)$, termed likelihood. According to Bayes' rule, the posterior equals the likelihood times the prior divided by the evidence:

$$\boldsymbol{Posterior} = \frac{\boldsymbol{likelihood\ x\ prior}}{\boldsymbol{evidence}}$$

This makes it possible to use Bayes' rule to transform the probability of the data given the hypothesis $P(D|H)$ to the probability of the hypothesis given the data ($P(H|D)$ which is most usually what scientists want to know.

Bayesian estimation. Bayesian inferential statistics are essentially just a reallocation of credibility based on collected data. Bayes' rule is used as the mathematical formula for this reallocation of belief. Our belief of where credibility is allocated serves as parameters in a mathematical model of the hypothesis. The first step in the analysis is therefore to develop a suitable model, which constitutes the parameters of the hypothesis. This depends on the research question, but typically involves the mean, differences of means between groups, standard deviation, differences in standard deviations of groups and so on. Bayes' rule then provides a rational mathematical rule about how prior knowledge should be updated in light of new data. Bayesian estimations is thereby a way of investigating the most credible values of parameters by updating current knowledge given new data.

Throughout history, the use of Bayesian estimation has been limited due to calculation difficulties, in particular integration over all model parameters to obtain an estimate of the evidence, $P(D)$. However, with the computational power and Markov Chain Monte Carlo (MCMC) technologies that are available today, it is possible to use Bayesian methods for statistical test regularly done in psychological research. The way this is done is by establishing priors which then can serve as a basis along with the collected data to generate posterior distributions with MCMC methods. A prior distribution is a distribution of credible parameter values that are based on previous knowledge before the data is collected and the posterior is the distribution of credible parameter values given the collected data. Parameter values that are consistent with the data becomes more credible than parameter values that are inconsistent with data. The prior distribution can be based on theory, previous findings or kept vague distributing credible values evenly across model parameters.

The choice of the prior distribution is an important aspect and should be a reflection of the research question. The prior should not just trivially presume a desired outcome without evidential support because it can affect the posterior distribution but, by and large, this is not problematic. In terms of Bayesian estimation, the prior is overwhelmed by the data reasonably quickly and different priors can be used to see if the choice of prior makes any difference to the results of the analysis (Kruschke, 2012; Rouder, 2018). An informative prior based on literature or the researcher's experience can result in an efficient model as long as the true effect is close to that prior, but will be inefficient (i.e., will not convincingly support or refute the hypothesis) if the prior is not close to the true effect. On the other hand, a less informative, vague prior that is centered on zero often requires a larger sample size but is more likely to pick up on a wider spectra of possible effect sizes (Schönbrodt & Wagenmakers, 2018). For example, suppose that the streets are wet, and we want to find out the cause of this. Our prior knowledge and experience might lead us to think that the wet streets were probably caused by rain because it is a rainy season. However, our prior knowledge can be specified more broadly and suggest several possible causes, such as washing of streets or a broken pipe. To investigate the most probable cause we might therefore collect data. If the humidity is high, the sky filled with clouds and we cannot see any street washing machines the probability that rain caused the wet streets is increased, because the collected data are more in line with that possibility than other possibilities. The alternative prior beliefs, are in this case, quickly overwhelmed by the collected metrological data.

Highest density intervals. The highest density intervals (HDI), also called highest posterior density (HPD) interval, is the range of the most credible values within the posterior distribution, and most usually an interval that spans 95% of the distribution (cf. Lindley 1965). The values inside the interval are more credible than values outside the interval and can be used in model or group comparison. A narrow HDI suggests high precision of the estimated parameter values. As more and more data are collected the HDI becomes increasingly narrow and more precise (Kruschke, 2015). Following this approach, the credibility of the Null value can be assessed by examining the posterior distribution in relation to where the Null value falls. Kruschke (2012) argues that by specifying a range of practical equivalence (ROPE) around the Null value it is possible to make a rational decision about the probability of the alternative and Null hypothesis in light of the data. If the HDI falls entirely within the ROPE, the Null value can

be accepted and if the HDI falls entirely outside the ROPE the Null hypothesis may be reasonably rejected. Bayesian estimation tells us exactly what we want to know – the probability of the hypothesis given the data. However, using Bayesian estimation for hypothesis testing is a somewhat controversial topic. Some argue on philosophical grounds (Rouder, Morey, Verhagen, Province, & Eric-Jan Wagenmakers, 2016) that it is only permissible to do hypothesis testing with likelihood ratios, and so the Bayes' factor should be used instead. These authors argue that if the Null model of no effect is considered important then both the Null and alternative model should be realized in the analysis to test for the categorical difference between null effects and effects.

Bayes' factor. The Bayes' factor (BF) is used to compare two models using Bayes' rule. The BF is the ratio of the probability of data given one hypothesis divided by the probability of the data given another competing hypothesis [i.e., $P(D|H_1) / P(D|H_0)$]. The BF estimates how many times more likely one model is compared to the other, and multiplied by a prior ratio [i.e., $P(H_1) / P(H_0)$] informs on which model we should believe in most given our current knowledge (Kruschke, 2012). Most usually the prior ratio, $P(H_1) / P(H_0)$, is set to 1, reflecting no prior knowledge about which hypothesis is true. The BF is applicable for testing an alternative hypothesis (H_1) against the Null hypothesis (H_0) by creating a model for the Null hypothesis with a high probability on zero. Conventionally, if the goal is to test the alternate hypothesis, H_1, the BF is denoted BF_{10}, whereas if the Null hypothesis, H_0, is of primarily focus BF_{10} is inverted (i.e., $1/BF_{10}$) and denoted BF_{01} (Wagenmakers, Lodewyckx, Kuriyal., & Grasman, 2010). Testing hypotheses with BFs are appropriate in situations where the only interest is to investigate if there is or is not an effect and not information about the precision of the estimated effect. The precision of an effect requires relevant summarization of the posterior distribution, most usually in terms of the HDI (Kruschke, 2012). Unlike Bayesian parameter estimation, a particular problem associated with BF is its sensitivity to the choice of likelihood prior distributions used to model the alternate hypothesis (Simmons, 2011; Sinhary & Stern, 2002). Due to this problem, there has been extensive work in developing, so called objective, default model distributions that can be used in large variety of applications. In this regard, Jeffreys-Zellner-Siow (JZS) priors are commonly used as default likelihood distributions (Jeffreys 1961; Liang, Paulo, Molina, Clyde, & Berger, 2008; Rouder & Morey, 2012; Zellner & Siow, 1980). JZS priors are Cauchy distributions drawn from the *t*-distribution with one degree of freedom (Liang, et al., 2008).

The width of the prior distribution used to model the likelihood for the alternate hypothesis, $P(D|H_1)$, expresses the plausibility of the existence of certain effect sizes. For instance, assuming a large effect size the prior distribution should be broad, reflecting the idea that strong evidence will be obtained in support of the alternate hypothesis if that hypothesis is true. Alternatively, if we assume a large effect size and so specify a broad model prior, but the true effect size is actually small, a large amount of data will be required to support the alternate hypothesis because the BF tends toward supporting the Null hypothesis with increased uncertainty.

Of most importance for sequential testing, Bayesian estimation and the Bayes' factor do not suffer with problems associated with mass significance testing using *p*-values, because Bayesian inference depends only on the likelihood of the data and prior knowledge of the analyst and not on precisely how many tests are conducted. In principle, sequential analysis is no problem for Bayesian analysts, because repeatedly testing the data using Bayesian methods does not change the interpretation of the data.

Sequential testing with BF

Schönbrodt et al., (2015) detail a procedure for sequentially testing data using the Bayes' factor, which they term sequential Bayes' factor (SBF) analysis. With SBF analysis, a Bayes' factor is computed as the data is being collected until it reaches a certain threshold (stopping for success) or unlikely to provide compelling support for one hypothesis or the other with all available resources (stopping for futility). The first step of the procedure is to decide on thresholds, in form of boundaries, to declare sufficient evidence for H_0 or H_1. This could for example be a BF ≥ 6. The next step is to choose a prior distribution of effect sizes to model the likelihood for H_1 [*i.e.*, $P(D|H_1)$]. Following these initial steps, the test is ready to run on a minimal number of participants in each group. If the BF does not cross the desired threshold the sample size is increased and the data tested again until the BF threshold is crossed, the researcher gives up, or a maximum number of participants is tested in line with the time and resources available.

Despite the promise of sequential Bayes' Factor (SBF) analysis and subsequent optional stopping of studies a number of potential problems remain. Early trajectories of BF, with small samples and imprecise estimates, tend to favor the Null hypothesis (Sanborn & Hills, 2014; Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, 2017). BFs are sensitive to the prior

distributions used to model the likelihood and judiciously or misguided specified prior distributions can increase the chance of obtaining BFs in desired direction (Sanborn & Hills, 2014). Moreover, BFs do not reveal the magnitude or certainty of an effect (Kruschke & Liddell, 2018). Sole reliance on the BF and incorporation of BF thresholds risks dichotomous hypothesis testing which, from the perspective of *p*-values, is considered a hazardous path for psychological science (see Cumming, 2014 for comprehensive discussion).

Sequential Bayesian parameter estimation that uses the HDI as an estimate of precision overcomes problems associated with SBF designs for efficiency (Schönbrodt et al., 2017; Kruschke, 2014). Bayesian parameter estimation is less sensitive to the priors which are quickly overwhelmed by the data (Kruschke, 2011, 2012). Moreover, stopping on the basis of precision is unaffected by the underlying value of the parameter, and so does not bias effect size estimates. Yet, sequential Bayesian parameter estimation for precision has only been addressed cursorily in the psychological literature (Kruschke, 2012). The downside of stopping for precision is that relatively large samples sizes are required. In an ideal world, scientists would like large sample sizes to obtain precise and stable parameter estimates but, in psychology, this is rarely possible due to time and resource limitations.

In the present paper, Monte Carlo simulations were performed to create hypothetical data with different effect sizes in order to test the efficiency of using sequential analysis with Bayesian inference. This was done in order to investigate if there are potential gains in efficiency by implementing sequential analysis in psychological research. The problems that are related to using NHST or solely BF in sequential analysis has been described throughout the introduction. The sequential analysis of simulated data in the present paper will therefore be performed with a method that combines BF and Bayesian estimation. The method is evaluated on the basis of succeeding or not in being an efficient and accurate method of detecting effects in data drawn from hypothetical populations.

Method

Monte Carlo simulations were performed in the statistical environment R, version 3.5.2 (R Core Team, 2018) to test the efficiency of sequential analyses with two different stopping rules defined by Bayesian parameter estimation in terms of the HDI as well as BFs with specified

boundaries. BF power analyses with fixed *n* were conducted as well in order to compare with the results from the sequential analyses.[2]

The effect size (δ), in terms of standardized differences in means, were altered between δ = .2, δ = .5 and δ = .8 which are considered to be a small, medium, large effect sizes respectively according to Cohen (1988). These effects are in the range of reported effects in psychology according to a meta-analysis done by Bakker, van Dijk and Wicherts (2012). They also report that the average effect size of meta-analyses in psychology is *d* =.5.

Simulations

The simulations presented here focus on testing the Null hypothesis of no difference between groups as compared to the alternate hypothesis of a difference between groups using sequential Bayes' factor *t* tests and Bayesian parameter estimation. In NHST, this is usually done using Student's (1908) *t* test. The *t* statistic is one of the most frequently used test statistics in psychology, and so adopting a Bayesian version of Student's t test to investigate the operating characteristics of sequential Bayesian analysis provides a widely applicable example.

The simulations were conducted by first generating two populations. Both populations comprised 1000000 units, and both populations were sampled from a normal distribution with a standard deviation (*SD*) = 1. One population was drawn from a normal distribution with a mean (*M*) = 0. The other population was drawn from a normal distribution with M = 0.2, 0.5, or 0.8, corresponding to small, medium and large effect sizes (δ) respectively.

Power analyses with fixed *n* were conducted to create a baseline in order to evaluate the potential gain in efficiency of using sequential testing. Statistical power was obtained by drawing two samples from defined populations. The 2 samples were always of equal size varying from *n* = 10 to *n* = 150. Every sample was randomly drawn from the hypothetically defined populations and tested 1000 times by way of a standard Bayes' factor *t* test (Rouder, Speckman, Sun, Morey, & Iverson, 2009) with a threshold of 3 and default prior of $\sqrt{2/2}$. This was done for all three δ. Power was defined as the percentage of simulations in which the BF was greater than or equal to 3 (BF ≥ 3).

Computational simulations for assessing the operating characteristics of sequential analyses were also tested using independent samples Bayes' factor *t* tests, as well as relevant

[2] The full r-code used for the simulations is provided in this link: https://github.com/pierreklintefors/MasterThesis

parameters estimated using Bayesian estimation. If one of the stopping rules: BF ≥ 3, or 95% HDI-width ≤ *SD**.50, were fulfilled the function stopped and the number of participants, *n*, was recorded for that simulation. If neither stopping rule was fulfilled, one participant was added to each group and both groups tested again. This procedure was repeated until one of the stopping rules was fulfilled or until a maximum sample size, $Max_n = 100$, was reached. This testing schedule was repeatedly undertaken 500 times for each of the three different effect sizes (δ = .2, δ = .5, δ =. 8). The maximum sample size, Max_n, was set to 100 in the simulations with sequential analyses to limit the time of every simulation due to restrictions in computational power (using a standard laptop computer, and parallel processing over all available cores, each simulation typically took 2-4 days to complete).

BF was calculated by using the package "BayesFactor" which includes a default BF *t* test (Morey & Rouder, 2018). The HDI was calculated using the package "BEST" (Kruschke & Meredith, 2018) along with the package "HDInterval" (Meredith & Kruschke, 2018). The number of MCMC steps to estimate the posterior distribution was set to 10000 with a burn-in of 2000. The burn-in steps refer to the initial portion of the chains that are discarded to avoid biasing the estimation because these steps tend to be unrepresentative of the posterior distribution (Kruschke, 2015). The package "snowfall" (Knaus, 2015) was used to optimize the code (i.e., decrease the time required for each simulation) by running parallel simulations over the available cores. However, in this case 'snowfall' could only be used for the BF analysis. The package "BEST" as used for Bayesian estimation calls JAGS (Plummer, 2003), which by default is optimized for parallel processing and so 'snowfall' does not make the simulation any faster in this case. All simulations were performed on a HP Pavilion 15 Notebook PC with 4 available cores.

Results

Figure 2 shows the result of the BF power analyses with fixed sample size, *n*, varied from $n = 10$ to $n = 150$, with small, medium, and large effect sizes, respectively. The power of the samples with small effect size, δ = .2, depicted in panel A of Figure 2, reached a maximum of 20% power when $Max_n = 150$ was reached. Panel B of Figure 2 shows that the simulations with medium effect size, δ = .5, requires $n \geq 90$ to have a power of at least 80% and $n > 118$ for a power of at least 90%. Panel C shows that a large effect size, δ = .8, requires $n > 33$ to obtain a power of at least 80%, $n > 42$ for at least 90% power and $n > 80$ for 100% power.

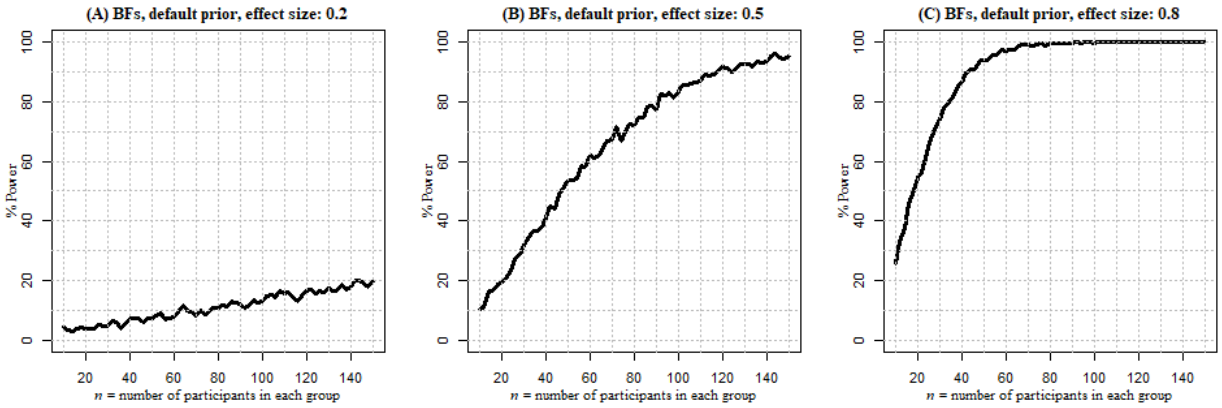

Figure. 2. Bayes' Factor (BF) Power, defined as: the proportion of 1000 simulations for different samples sizes (*n*), varying from *n* = 10 to *n* = 150, that reached a BF ≥ 3. The samples were tested with a BF *t* test with a default prior of $\sqrt{(2)}/2$, recognized as a 'medium' prior scale, as recommended by Morey et al. (2011) and Rouder, Morey, Speckman, and Province, (2012). Panel A shows the power for samples with small effect size (δ = .2), panel B shows the power for samples with medium effect size (δ = .5) and panel C shows the power for samples with large effect size (δ = .8).

Figures 3 and 4 shows the results from the simulations with the sequential analyses that includes Bayesian estimation with a 95% HDI-width. Figure 3 shows the proportion of stopped simulation when δ = 0 (i.e. Null hypothesis was true), and is defined as false alarms. The false alarm rate reached just 8.8% by the time the maximum sample size, Max_n =100, was reached. There was a small proportion of the false alarms (< 5%) that occurred at the minimal sample size, Min_n = 15, which increased gently in rate until *n* > 40 where the false alarm rate starts to stabilize at about 8%. This contrasts sharply with the false alarm rate (i.e., Type I error rate) for sequential frequentist NHST as shown in Figure 1, indicative of a 40% false alarm rate with a sample size of *n* = 100 in each group.

Figure 4 shows the proportion of simulations with effect sizes of δ = .2, δ = .5, and δ = .8 that stopped as a result of detecting the effect based on the stopping rules together with the sample sizes of the stopped simulation. The proportioned of stopped simulations are labeled as the success rate of the simulations because they were correctly stopped based on an existing effect. As shown in Panel A, the success rate of detecting the small effect size, δ = .2, increased together with sample size and reached a maximum of 30% at the maximum sample size, Max_n

=100. Panel B shows the success rate of the simulations with a medium effect size, δ = .5, which was considerably higher than for the simulations with a small effect size. The success rate climbed above chance (> 50%) when $n > 40$. When $n > 80$ the success rate of the simulations was around 85% and around 90% of the simulations were stopped for success when the maximum sample size, Max_n, was reached. The success rate for the simulations with δ = .8, as shown in Panel C, was the highest. The simulations with δ = .8 reached a success rate of 85% when $n > 33$, 90% when $n > 38$ and 100% when $n \geq 72$. So, when the effect size was large, δ = .8, the simulations never reached Max_n but always stopped according to the stopping rules at a maximum n =72.

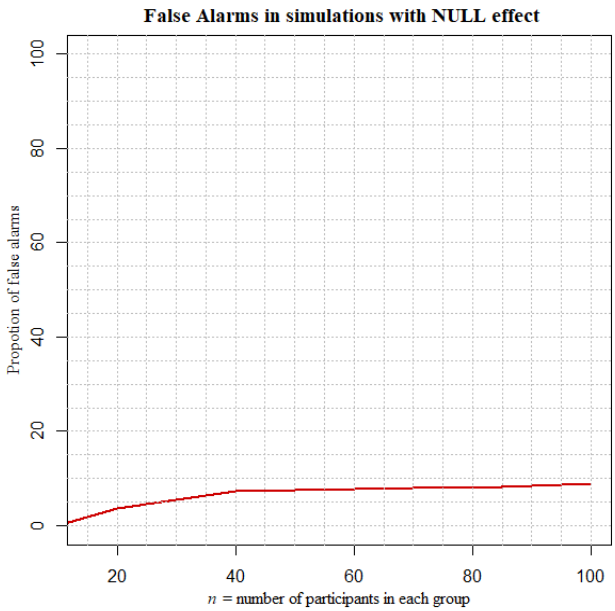

Figure. 3. The proportioned stops of fulfilling stopping rule Bayes' Factor (BF) ≥ 3 when effect size was Null (δ = 0), termed as false alarm rate. The simulations tested for a group difference when the Null hypothesis was true using sequential BF *t* tests after every added participant until Max_n=100 was reached.

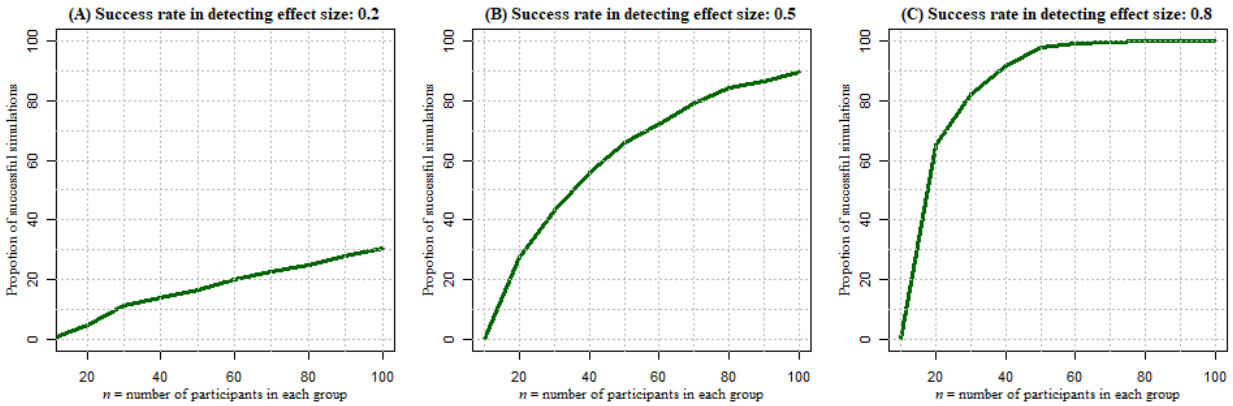

Figure. 4. The proportion of simulations that stopped according to the stopping rules Bayes' Factor (BF) ≥ 3, and 95% HDI ≤ *SD**.50. The simulations tested for a group difference using sequential BF *t* tests after every added participant until Max_n = 100 was reached. Panel A shows the proportion of simulations stopped for success with a small population effect size (δ = .2), panel B shows the proportion of simulations stopped for success with a medium population effect size (δ = .5) and panel C shows the proportion of simulations stopped for success with a large effect size (δ = .8).

The specified proportions of the stopped simulations as a result of fulfilling the stopping rules are depicted in Table 1. The HDI-width rarely came close to the stopping rule of ≤ *SD**.50. The terminated simulations stopped almost solely on the account of the stopping rule BF ≥ 3 or reaching Max_n. A new set of independent simulations with δ = .2, δ = .5 and δ = .8 was run to investigate the required sample size to reduce the HDI-width to the stopping rule of *SD**.50. Figure 5 shows the simulated HDI-width for a range of different sample sizes (*n* = 10-500). The simulations were done with effect sizes of δ = .2, δ = .5, and δ = .8 which are presented in panel A, B and C, respectively. The effect size did not affect the HDI-width which can be seen by comparing the panels: A, B and C of Figure 5, which only differs marginally. The HDI-width was not less than *SD**.50 until *n* > 120 which exceeded the Max_n of the earlier simulations with sequential simulations.

Table 1. Specified proportion of stopped simulations for different effect sizes (δ)

δ	Stopped simulations†	Bayes' Factor ≥ 3	95% HDI-width ≤ *SD**.50
0	10.6%	8.8%‡	1.8%
0.2	30.6%	29.4%	1.2%
0.5	89.4%	89.2%	0.2%
0.8	100%	100%	

† The combined proportioned simulations as a result of fulfilling any of the two stopping rules before reaching Max$_n$.

‡ When δ = 0, the proportion of stops resulted by the fulfilling of stopping rule: Bayes' Factor ≥ 3 accounts as false alarms.

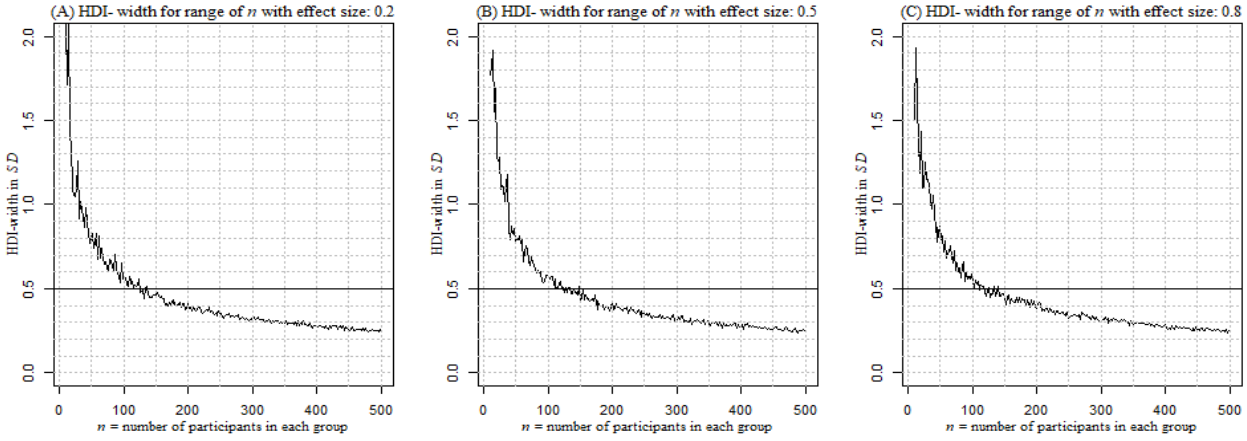

Figure. 5. Simulations of HDI-width for a range of different sample sizes (*n*) ranging from 10-500. Panel A, B and C shows simulation with a small (δ = .2), medium (δ = .5) and large (δ = .8) effect, respectively. Sample size affects the HDI-width but effect size does not have a substantial effect.

In order to investigate the suitability of using a BF boundary of 3 as a stopping rule, another round of simulations with sequential analyses were conducted but with increased BF boundary set to 6 instead of 3. An aim here was to test the effects of increasing the BF bound, so Bayesian posterior estimation and subsequent assessment of HDIs was excluded from this round of simulations, to reduce the time of the simulations. Figure 6 shows the false alarms of the sequential analysis simulations with the BF bound set to 6, under conditions in which the Null hypothesis was defined as True, δ = 0. The false alarm rate reached its maximum of 5% when all

sample sizes were included, which is generally considered acceptable for psychological research. Figure 7 shows the success rate of detecting effects with the new stooping rule of BF ≥ 6. Panel A shows the simulations with δ = .2 where the success rate reached a maximum of 20% at Max$_n$. Panel B shows the simulations with δ = .5 with a maximum success rate of 80% when n > 95. Panel C shows the simulations with δ = .8 where the success rate reached 85% when n > 39, 90% when n > 48 and 99-100% when n > 90.

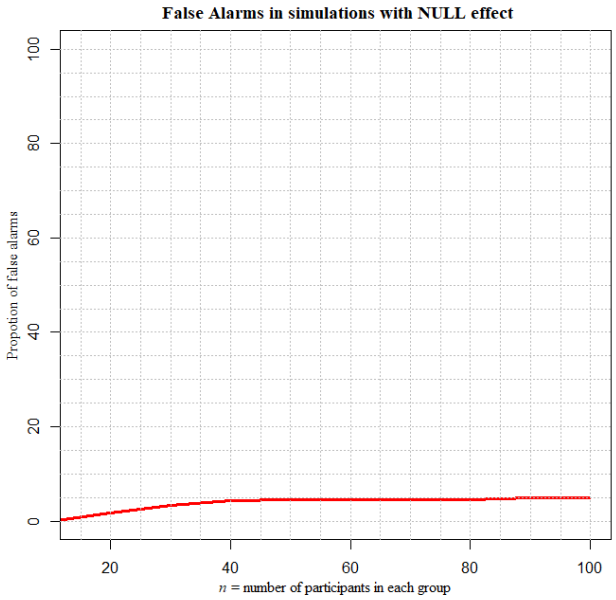

Figure. 6. The proportion of stops fulfilling the stopping rule Bayes' Factor (BF) ≥ 6 when effect size was Null (δ = 0), termed as False alarm rate. The simulations tested for a group difference when the Null hypothesis was true using sequential BF *t* tests after every added participant until Max$_n$=100 was reached.

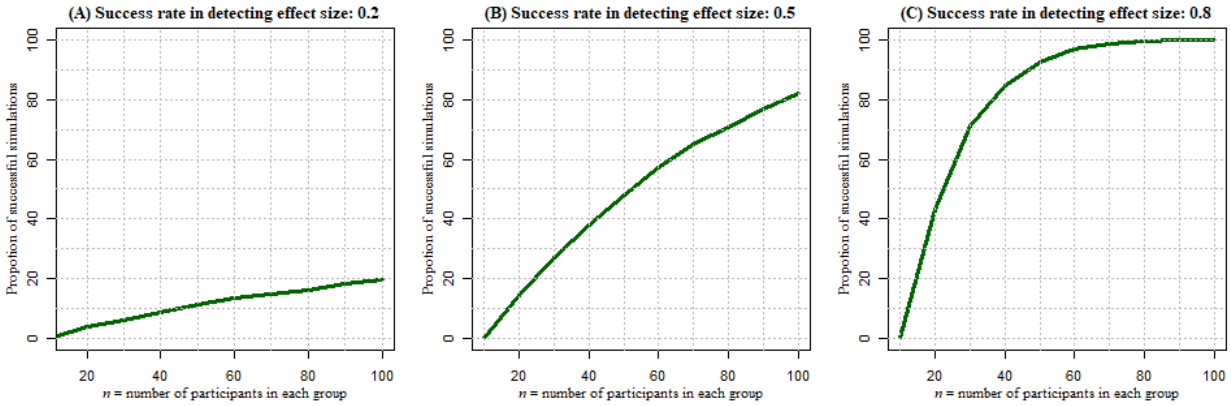

Figure. 7. The proportion of simulations that stopped according to the stopping rule Bayes' Factor (BF) ≥ 6. The simulations tested for a group difference using sequential BF *t* tests after every added participant until Max_n=100 was reached. Panel A shows proportion of simulations stopped for success with a small population effect size (δ = .2), panel B shows proportion of simulations stopped for success with a medium population effect size (δ = .5) and panel C shows proportion of simulations stopped for success with a large effect size (δ = .8).

Discussion

The stopping rules of BF ≥ 3 and HDI-width ≤ *SD**.50 were tested on simulations with sequential *t* tests with δ = .2, δ = .5 and δ = .8. The simulations with the sequential testing on δ = .2 had a success rate of 30% when all sample sizes up till Max_n=100 were included. This 30% success rate exceeded the statistical power of 20% obtained with a fixed *n* =150. The sequential analysis design, therefore, improved the success rate of detecting the effect with a smaller sample size. However, 30% is still under chance level (i.e., 50%). This is in line with earlier findings where δ = .2 in a BF analysis with default priors was predicted to require a large sample size of between 200-300 participants (Stefan, Gronau, Schönbrodt, Wagenmakers, 2019).

An alternative to improve the success rate and keep the sample size low is to lower the BF boundary, which will make it easier to stop the sequence early. However, this will result in an increased risk for false alarms in populations with small or no effect size, and may potentially bias estimates of the true effect size, because precision in terms of narrow HDIs require reasonably large sample sizes (Kruschke, 2015; Perugini, Gallucci, & Costantini, 2014). It is also possible to change the priors used to model the alternative hypothesis from default priors to more informed priors. BFs are sensitive to the width of the prior used to model the likelihood and the

use of an informed prior may result in a more effective sequential testing procedure than habitual use of a default prior. This approach may prove useful for studies examining special populations (such as people suffering from a rare illness) in which it may be very difficult, if not impossible, to recruit a sufficiently large number of participants. However, an informed prior increases the risk of missing a wider spectrum of potential effects and can increases false alarms. Stefan et al. (2019) has conducted simulations that show that informed priors generally require a lower sample size but are also more prone to false alarms for small sample sizes. To lower the stringency of the stopping rules might therefore be a bad idea. If a small effect is expected or considered to be valuable to investigate, this should ideally be expressed in a high Max_n, and not by easily fulfilled stopping rules.

Meta-analyses show that many of the studies in psychology with small effect sizes are underpowered which results in low replicability (Bakker et al., 2012). This is due to the small sample sizes which causes offsets and uncertainty in the accuracy of the estimated effects (Perugini et al., 2014). This is problematic because the results from these underpowered studies are generally unreliable, reducing confidence in psychological science. Improving efficiency is important, and it is the foundation for the present paper, but it should always be done in a proper manner that does not sacrifice the accuracy of the estimation and the validity of the study. If there are expectations of potentially small effect sizes that are considered to be valuable to investigate, the set maximum sample size should reflect this fact by being large enough for the study to obtain accurate estimations. On the other hand, if small effects are not of interest, sequential testing is a good way of terminating the study early to save resources.

Even if the stopping rules only had a 30% success rate for the small effect size, they were more successful in detecting the larger effect sizes, $\delta = .5$ and $\delta = .8$. For $\delta = .5$, the success rate was 90% when all sample sizes were included up till $Max_n = 100$; respectively, 100% for $\delta = .8$ when $n \geq 72$. These success rates have smaller sample sizes compared to the equivalent obtained statistical power from the BF power analyses that presented in Figure 2. When n was fixed and $\delta = .5$, statistical power analyses showed that at least 120 participants are required in each group ($n > 120$) to obtain 90% power. Respectively, when $\delta = .8$, statistical power of 100% requires $n > 80$. The simulations with sequential analyses detected the effects with smaller sample sizes and therefore are more efficient than the fixed n design.

Increasing the BF boundary from BF ≥ 3 to BF ≥ 6, in the present paper, affected both the false alarm rate and efficiency of the sequential testing procedure. This is in line with previous simulations reported in the literature (Schönbrodt, Wagenmakers, 2018). With increased BF boundary from 3 to 6 the false alarm rate dropped by 3.8 percentage points. Efficiency, however, in terms of the required sample sizes for successfully detecting an effect, was affected as well. For δ =.2 the maximum success rate of the simulations with the stopping rule of BF ≥ 6 was 20% - a 10 percentage point drop as compared to the simulations with a stopping rule of BF ≥ 3. When δ =.5, the stopping rule of BF ≥ 6 had a success rate of 80% when $n < 90$, which is 22 participants more than required for the same success rate with the stopping rule of BF ≥ 3. When δ =.8, the success rate was 90% when $n > 48$ and 99% when $n > 90$, which were 10 respectively 18 more participants than required for the same success rates with the stopping rule of BF ≥ 3. Nonetheless, the false alarm rate with a stopping rule of BF ≥ 3 was 8.8%, which exceeds the normally accepted *alpha* level of 5% in psychological research (Fisher, 1925). A BF boundary of 6 may therefore be a more suitable stopping rule for the sequential analyses. The loss in efficiency, in terms of a larger sample size required to detect an effect, might be needed in order to keep the false alarm rate low. To investigate the potential gain of using the stopping rule of BF ≥ 6, for a sequential analysis design, further BF power analyses with fixed *n,* that implements this stopping rule, are needed for comparison.

Larger effect sizes will tend to be detected earlier than small effect sizes in sequential designs with a stopping rule based on crossing a BF boundary and subsequently result in smaller *n* but wider HDI's (Kruschke, 2012; Schönbrodt & Wagenmakers, 2018). Inspection of false alarms shown in Figure 3, however, shows that large BF-values and early stops does not necessarily mean that there is a large true effect in the population. When there were true effects in the populations, a small proportion of the stopped simulations can have a high BF-values for $Min_n = 15$, especially for δ = .5 and δ = .8. These high BFs tend to happen early with small sample sizes and could be a result of a lucky sequence of draws. High BF-values obtained early in the sequential testing should therefore be interpreted with caution. According to simulations done by Stefan et al. (2019), the risk of false alarms decreases with an increased BF boundary. However, when the boundary was higher than 10 there was only a slight improvement of the error rates by increasing the boundary. It is important to not let this property lead to misleading evidence which can be done by comparing the *n* of these early stops to *n* of what to be expected

for certain effect sizes as well as the precision of the estimation. So, this suggests how important it is not to only rely on BF because it can have misleading effect if the precision of estimation, in terms of HDI-width, is ignored.

Another approach for model testing is to use Bayesian estimation with a region of practical equivalence (ROPE) rather than Bayes' factor. This approach is advocated by Kruschke (2012, 2015), but has otherwise received little discussion in psychological literature (but see Rouder, et al., 2018). By assessing the HDI and defining a ROPE - an interval of effect sizes that are considered to be practically equivalent with zero - it is possible to discard small effects that are negligible, for example an δ in the range of -.1 to .1. A ROPE can also be useful in cases where a specific point Nill-Null effect is improbable. This can be the case in cognitive psychology when investigating correlations between cognitive domain, such as memory, and motivational behavior. These correlations can occur via vast spectrum of neurological mechanisms or third variables. It is not reasonable to expect that an exact zero relationship is probable in such cases. In this case, the inference is based on the posterior distributions of key parameters. The advantage of an easy computational structure to test a point Null value using a Bayes' Factor approach is therefore less obvious in these situations. In terms of Bayesian estimation, and continuous probability distributions, there is no probable specific point of Nill-Null and it might not be more beneficial than Bayesian estimation (Williams, Bååth, & Philipp, 2017). However, it is generally not recommended for drawing conclusions about the presence or absence of effect with Bayesian estimation. If the goal of a study is to simply determine whether an effect exists or not, the BF approach is generally favored (Rouder, et al., 2018, Wagenmakers, et al., 2019).

A potential way of unifying the Bayes' factor approach and Bayesian estimation of posterior intervals is to use so called "spike-and-slab-priors" (George & McCulloch, 1993; Mitchell & Beauchamp, 1988; Rouder, Haaf, & Vandekerckhove, 2018; Williams, Bååth, & Philipp, 2017). This approach provides for a combination of a likelihood ratio test of a Null point hypothesis as compared to the alternative hypothesis and posterior parameter estimation of the effect. The total mass of prior probability is divided between a spike representing the Null hypothesis and slab representing the alternative hypothesis. The spike prior refers to the part of the mass that is placed on the point of zero and the slab is the distribution of the prior representing the effect.

The spike-and-slab-model can be interpreted as a hierarchical relationship between prior-comparison with Bayes' Factor and Bayesian estimation with HDI and ROPE. By this interpretation, the different levels of the same model based of Bayesian inference. This interpretation is implemented by Kruschke's (2012) approach to test hypothesis with HDI and ROPE. With an HDI and ROPE-decision rule, the focus lies on continuous parameter values of the posterior distribution of the most probable hypothesis in the comparison. Bayes' Factor focuses on that higher level of the model. Namely, the decision of the most probable hypothesis. Say that the alternative hypothesis is favored by Bayes' Factor, in relationship to the prior distribution, to be more probable than the Null hypothesis. The HDI and ROPE-decision rule is based on the relationship between the sub-interval of posterior distribution of the most likely hypothesis which in this case would be the alternative hypothesis (Kruschke & Liddell, 2018). In other words, a focus on estimation of a posterior distribution can be seen as parameter estimation after choosing the slab (effect distribution) of one hypothesis. This interpretation is problematized by Rouder et al. (2018) because of a lack of underlying reasons to focus the slab on estimation. Nonetheless, even if there is a certain ambiguity of the interpretations of spike-and-slab-priors, they are being depicted in the literature as a good way of combining two Bayesian approaches.

The attempt to combine Bayes' factor and Bayesian estimation in the present paper was done by including both BF boundary and HDI in the decision rules to determine the presence of an effect. However, choosing the limits of a suitable HDI-width can be problematic. The stopping rule of an HDI-width ≤ *SD**.50 used in the sequential testing simulations in the present paper were rarely fulfilled. The independent HDI-width simulations for different samples sizes that are plotted in the bottom right panel of Figure 5 shows that a width of *SD**.50 might not be relevant for simulations with a $Max_n \leq 114$. The indication is that an HDI ≤ *SD**.50 is too narrow to serve as a suitable stopping rule for sample sizes under 100. However, the simulation of HDI-width of different sample sizes was just performed once for every effect size and should had been repeated for more precise estimation of the required sample size to reach *SD**.50. If the criterion of the stopping rule is increased from a width of *SD**.50 to a width of *SD**.60 it could potentially improve the efficiency without sacrificing too much of the accuracy.

Figure 8 shows the success rate from Figure 4 combined with the corresponding HDI-width from Figure 5. Panel A constitutes the results from the simulations with δ = .5 which is

considered to be the average reported effect size in psychology (Bakker et al., 2012). An HDI-width ≤ *SD**.60 corresponding with $n > 80 \leq 100$ and a success rate > 85% might therefore make it a more suitable stopping rule. It would reduce the number of simulations that reached Max_n and potentially increase the number of stopped simulations. However, to test the latter would require another simulation with the new stopping rule implemented. Regarding the simulations with δ = .8, it is difficult to get a narrow HDI with a large effect size because all the simulations stopped when $n \leq 72$. Panel B of Figure 6 includes the success rate and HDI-width of the simulations with δ = .8. At the point of the 100% success rate the HDI-width was around *SD**.70 and a narrower HDI-width would require a larger sample size.

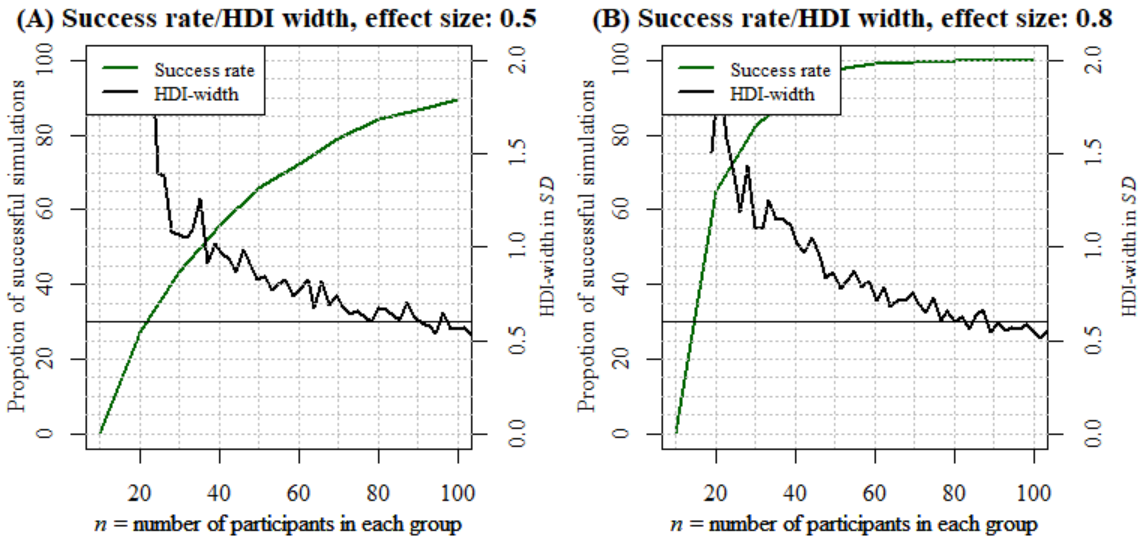

Figure. 8. The proportion of sequential analysis procedure that stopped according to the stopping rules Bayes' Factor (BF) ≥ 3, and 95% HDI ≤ *SD**.50. The simulations tested for a group difference using sequential BF *t* tests after every added participant until Max_n=100 was reached. HDI-width in *SD* for the corresponding *n* is also plotted together with a black horizontal line to show the limit for a potential stopping rule of 95% HDI ≤ *SD**.60. Panel A shows the success rate and HDI-with for simulations with medium effect size (δ = .5.). Panel B shows the success rate and HDI-with for simulations with large effect size (δ = .8),

Limitations

The simulations in the present paper were Monte Carlo with randomly generated units that followed a normal distribution. Simulated studies are specific to the conditions under

investigation. Without mathematical proof it is not possible to draw general statistical conclusions that hold beyond the simulations. Real data from actual experiments tend to be less predictable and so further work should be conducted with real data to gain an increased understanding of the operating characteristics of sequential Bayesian analyses. Due to limitations of time and computational power, the Max_n was set to 100 and the number of simulations set to 500. This limits the simulations suitability for detecting small effects as well as the precision of the estimations. To increase both these factors along with several comparable alternatives of BF-boundaries of 6, and higher, would strengthen the results.

Conclusion

The stopping rules of a BF ≥ 3 and 95% HDI-width ≤ *SD**.50 were more successful in detecting a medium and large effect, δ =.5 and δ = .8, than a small effect, δ = .2, with maximum sample sizes of 100 units in each group. If smaller effect sizes are to be accurately estimated, bigger sample sizes are required. BF *t* test is a suitable tool for sequentially testing the existence of an effect based on a likelihood ratio test. A BF boundary of 6 may be a more suitable stopping rule than 3 because it keeps the false alarms rate within an acceptable level. The complementary rule based on a 95% HDI-width is a good way of avoiding biased estimations caused by chance random draws in the sampling process. The rule's defined limits of HDI-width could possibly be increased from ≤ *SD**.50 to ≤ *SD**.60 if $Max_n \leq 100$. Along with the present paper, there are several advocates in the literature (Rouder, et al., 2018; Williams, et al., 2017) for a unity of the Bayes' factor approach for efficiency and Bayesian estimation approach for precision. Further work with data from real experiments is recommended to investigate fully the feasibility and potential gains of implementing sequential testing using BFs and Bayesian estimation in psychology.

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. doi:10.2307/2343787.

Bakker, M., van Dijk, A., Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, 7, 543–554. doi:10.1177/1745691612459060

Berry, S. M., Carlin, B. P, Lee, J. J., & Müller, P. (2010*). Bayesian Adaptive Methods in Clinical Trials*. Bosa Roca, United States: CRC Press Inc. doi: 10.1201/EBK1439825488

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. New York, NY: Routledge Academic.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*. 25(3), 7–29. doi:10.1177/0956797613504966

Debnath, L. & Basu, K. (2015) A short history of probability theory and its applications, *International Journal of Mathematical Education in Science and Technology*, 46(1), 13-39, doi: 10.1080/0020739X.2014.936975

Ellis, P. D. (2010). T*he Essential Guide to Effect Sizes: An Introduction to Statistical Power, Meta-Analysis and the Interpretation of Research Results*. Cambridge: Cambridge University Press.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*, Edinburgh: Oliver and Boyd.

George, E. I. & McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, 88(423), 881–889. doi:10.1080/01621459.1993.10476353

Jeffreys, H. 1961. Theory of probability (3rd ed.). New York: Oxford University Press.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science*, 23(5), 524–532. doi: 10.1177/0956797611430953

Koch, K-R. (2007) Probability. In: *Introduction to Bayesian Statistics.* Springer-Verlag Berlin Heidelberg

Kruschke, J. K. (2011). Bayesian Assessment of Null Values Via Parameter Estimation and Model Comparison. *Perspectives on Psychological Science*, 6(3), 299–312. doi: 10.1177/1745691611406925

Kruschke, J., K. (2012). Bayesian Estimation Supersedes the *t* Test. *Journal of Experimental*, 142(2), 573– 603. doi: 10.1037/a0029146

Kruschke, J., K. (2015). *Doing Bayesian Data Analysis* (2nd ed.). Amsterdam: Academic press.

Kruschke, J., K., & Liddell, T.M. (2018). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, 25(1), 178-206. doi:10.3758/s13423-016-1221-4

Kruschke, J K. & Meredith, M. (2018). BEST: Bayesian Estimation Supersedes the t-Test. R package version 0.5.1. https://CRAN.R-project.org/package=BEST

Lai, T., L. (1973). Optimal stopping and sequential tests which minimize the maximum expected sample size. *The Annals of Statistics*, 1(4), 659-673. doi:10.1214/aos/1176342461.

Laplace, P. S. (1774). Mémoire sur la probabilité des causes par les événements. *Mémoires de l'Académie royale des sciences de Paris (Savants étrangers)*

Lakens, D. (2014). Performing High-Powered Studies Efficiently With Sequential Analyses. *European Journal of Social Psychology*, 44(7), 701-710. doi:10.1002/ejsp.2023.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. J*ournal of the American Statistical Association*, 103, 410-423. Doi: 10.1198/016214507000001337

Lindley, D. V. (1965). *Introduction to probability and statistics from a Bayesian point of view, part 2: inference.* Cambridge: Cambridge University Press.

Meredith, M, & Kruschke, J. K. (2018). HDInterval: Highest (Posterior) Density Intervals. R package version 0.2.0. https://CRAN.R-project.org/package=HDInterval.

Mitchell, T. J. & Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404), 1023–1032. doi:10.1080/01621459.1988.10478694

Morey, R. D. & Rouder, J. N. (2011). Bayes' factor approaches for testing interval null hypotheses. *Psychological Methods*, 16(4), 406-419. doi: 10.1037/a0024377

Morey, R. D. & Rouder, J. N. (2018). BayesFactor: Computation of Bayes Factors for Common Designs. R package version 0.9.12-4.2. https://CRAN.R-project.org/package=BayesFactor

Neyman, J. & Pearson, E. S. (1933). On the Problem of the Most Efficient Tests of Statistical Hypotheses, *Philosophical Transactions of the Royal Society of London*, 231, 289-337. doi: 10.1098/rsta.1933.0009

Perugini, M., Gallucci, M., & Costantini, G. (2014). Safeguard power as a protection against imprecise power estimates. *Perspectives on Psychological Science*, 9(3), 319–332. doi:10.1177/1745691614528519

Proschan, M. A., Lan, K. K. G., & Wittes, J. T. (2006). *Statistical monitoring of clinical trials: a unified approach*. New York: Springer, 2006.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rouder, J. N., Haaf, J. M., & Vandekerckhove, J. (2018). Bayesian Inference in Psychology, Part IV: Parameter estimation and Bayes factors. *Psychonomic Bulletin & Review*, 25(1), 102–113. doi: 10.3758/s13423-017-1420-7

Rouder, J. N., & Morey, R. D. (2012). Default Bayes factors for model selection in regression. *Multivariate Behavioral Research*, 47: 877-903. doi:10.1080/00273171.2012.734737

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, 56(5), 356-374. doi: 10.1016/j.jmp.2012.08.001

Rouder, J. N., Morey R. D., Verhagen J., Province J. M. & Wagenmakers E-J. (2016). Is There a Free Lunch In Inference? *Topics in Cognitive Science, 8*, 520-547. doi:10.1111/tops.12214.

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. doi:10.3758/PBR.16.2.225.

Sanborn, A. N. & Hills, T. T. (2014). The frequentist implications of optional stopping on Bayesian hypothesis tests. *Psychonomic Bulletin & Review*, 21(2), 283-300. doi: 10.3758/s13423-013-0518-9.

Schönbrodt, F. D. & Wagenmakers, E-J. (2018). Bayes factor design analysis: Planning for

compelling evidence. *Psychonomic Bulletin & Review*, 25(1), 128-142. doi:10.3758/s13423-017-1230-y

Schönbrodt, F., D, Wagenmakers, E-J., Zehetleitner, M., & Perugini, M. (2017) Sequential Hypothesis Testing With Bayes Factors: Efficiently Testing Mean Differences. *Psychological Methods*, 22(2), 322–339. doi:10.1037/met0000061.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science,* 22, 1359-1366. doi: 10.1177/0956797611417632

Sinharay, S., & Stern, H. S. (2002). On the sensitivity of Bayes factors to the prior distributions. *The American Statistician,* 56, 196–201. doi:10.1198/000313002137

Stefan, A. M., Gronau, Q. F., Schönbrodt, F. D., & Wagenmakers, E-J. (2019) A tutorial on Bayes Factor Design Analysis using an informed prior. *Behaviour Research Methods,* 51(1), 1-17. doi:10.3758/s13428-018-01189-8.

Student. (1908). The probable error of a mean. Biometrika, 6, 1–25.

Wagenmakers, E-J., Lee, M., Lodewyckx, T., & Iverson, G J. (2008) Bayesian Versus Frequentist

Inference. In: Hoijtink H., Klugkist I., Boelen P.A. (Eds) *Bayesian Evaluation of Informative Hypotheses*. Statistics for Social and Behavioral Sciences. New York: Springer.

Wagenmakers, E.-J., Lodewyckx, T., Kuriyal, H., & Grasman, R. (2010). Bayesian hypothesis testing for psychologists: A tutorial on the Savage-Dickey method. *Cognitive Psychology,* 60(3), 158–189. doi:10.1016/j.cogpsych.2009.12.001.

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, A. J., Love, J., Selker, R., . . . Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review, 25,* 58-76. doi:10.3758/s13423-017-1343-3

Williams, M. N., Bååth, R. A., & Philipp, M. C. (2017). Using Bayes factors to test hypotheses in developmental research. *Research in Human Development,* 14(4), 321-337. doi:10.1080/15427609.2017.1370964

Yu, E. C., Sprenger, A. M., Thomas, R. P., & Dougherty, M. R. (2014). When decision heuristics

and science collide. *Psychonomic Bulletin & Review*, 21(2), 268-82. doi:10.3758/s13423-013-0495-z.

Zellner, A., & Siow, A. (1980). *Posterior odds ratios for selected regression hypotheses*. In J. M. Bernardo, M. H. De Groot, D. V. Lindley, & A. F. M Smith (Eds.), Bayesian statistics: 558-603. Valencia, Spain: University Press.