

# Controlling Decision Errors With Minimal Costs: The Sequential Probability Ratio $t$ Test

Martin Schnuerch and Edgar Erdfelder  
University of Mannheim

## Abstract

For several years, the public debate in psychological science has been dominated by what is referred to as the *reproducibility crisis*. This crisis has, inter alia, drawn attention to the need for proper control of statistical decision errors in testing psychological hypotheses. However, conventional methods of error probability control often require fairly large samples. Sequential statistical tests provide an attractive alternative: They can be applied repeatedly during the sampling process and terminate whenever there is sufficient evidence in the data for one of the hypotheses of interest. Thus, sequential tests may substantially reduce the required sample size without compromising predefined error probabilities. Herein, we discuss the most efficient sequential design, the sequential probability ratio test (SPRT), and show how it is easily implemented for a 2-sample  $t$  test using standard statistical software. We demonstrate, by means of simulations, that the SPRT not only reliably controls error probabilities but also typically requires substantially smaller samples than standard  $t$  tests and other common sequential designs. Moreover, we investigate the robustness of the SPRT against violations of its assumptions. Finally, we illustrate the sequential  $t$  test by applying it to an empirical example and provide recommendations on how psychologists can employ it in their own research to benefit from its desirable properties.

## Translational Abstract



Fostered by a series of unsuccessful attempts to replicate seemingly well-established empirical results, the reproducibility crisis has dominated the public debate in psychological science for several years. Apart from increasing awareness for the consequences of questionable research practices, the crisis has drawn attention to the shortcomings of currently dominating statistical procedures. Critically, conventional methods that allow for control of both Type I and Type II statistical error probabilities— $\alpha$  and  $\beta$ , respectively—often require sample sizes much larger than typically employed. Therefore, we promote an alternative that requires substantially smaller sample sizes on average while still controlling error probabilities: sequential analysis. Unlike conventional tests, sequential tests are designed to be applied repeatedly during the sampling process and terminate as soon as there is sufficient evidence for one of the hypotheses of interest. Herein, we discuss the most efficient sequential design, the sequential probability ratio test (SPRT), and show how it is easily implemented for the common  $t$  test to compare means of 2 independent groups. We demonstrate by means of simulations that the SPRT reliably controls error probabilities and requires smaller samples than standard  $t$  tests or other common sequential designs. Moreover, we investigate the robustness of the SPRT against violations of its assumptions. Finally, we illustrate the sequential  $t$  test by applying it to an empirical example and provide concrete recommendations on how psychologists can employ it in their own research to benefit from its desirable properties.

**Keywords:** hypothesis testing, efficiency, statistical error probabilities, sequential analysis, sequential probability ratio test

Critical tests of theories and hypotheses are at the heart of psychological science. A good theory makes clear-cut predictions that can be evaluated empirically, for example, in an experiment. Empirical tests of such predictions often take the form of binary

decisions: Based on the data, do we accept the hypothesis of interest or do we reject it, thereby corroborating or refuting the underlying theory? The most common statistical procedure in psychology to decide between conflicting hypotheses is usually

This article was published Online First September 9, 2019.

 Martin Schnuerch and  Edgar Erdfelder, Department of Psychology, School of Social Sciences, University of Mannheim.

This research was supported by a grant from the Deutsche Forschungsgemeinschaft (DFG, GRK 2277) to the Research Training Group “Statistical Modeling in Psychology” (SMiP). Parts of this article were presented at the 60th Conference for Experimental Psychologists (2018) in Marburg,

Germany. The authors thank Daniel W. Heck for helpful discussions and comments on an earlier version of the manuscript.

Correspondence concerning this article should be addressed to Martin Schnuerch or Edgar Erdfelder, Department of Psychology, School of Social Sciences, University of Mannheim, Schloss, 68131 Mannheim, Germany. E-mail: [martin.schnuerch@psychologie.uni-mannheim.de](mailto:martin.schnuerch@psychologie.uni-mannheim.de) or [erdfelder@psychologie.uni-mannheim.de](mailto:erdfelder@psychologie.uni-mannheim.de)

referred to as *null-hypothesis significance testing* (NHST). NHST has been harshly criticized in the past, and rightly so, as it is an inconsistent hybrid between two seemingly similar but, in fact, substantially different statistical theories: the theory of significance testing proposed by Fisher and the theory of statistical decision making by Neyman and Pearson (e.g., Bakan, 1966; Berger, 2003; Breidenkamp, 1972; Cumming, 2014; Dienes, 2011; Gelman, 2016; Gigerenzer, 1993, 2004; Goodman, 1993; Sedlmeier, 1996; Wagenmakers, 2007). Notwithstanding these criticisms, NHST has been the dominant procedure in behavioral science for decades. However, fostered by the reproducibility crisis in psychology (Asendorpf et al., 2013; Earp & Trafimow, 2015; Maxwell, Lau, & Howard, 2015; Open Science Collaboration, 2015; Pashler & Wagenmakers, 2012; but see Gilbert, King, Pettigrew, & Wilson, 2016), there is an increasing awareness of the pitfalls of NHST and the importance of rigorous control of decision errors in hypothesis testing.

According to Neyman and Pearson (1933), two types of errors can occur when deciding between a null hypothesis ( $\mathcal{H}_0$ ) and an alternative hypothesis ( $\mathcal{H}_1$ ): The null hypothesis is rejected when it is true (Type I error), or it is accepted when it is false (Type II error). By convention, the probabilities of Type I and II errors are denoted by  $\alpha$  and  $\beta$ , respectively. The complement of  $\beta$ ,  $1 - \beta$ , is referred to as the statistical power of the test. As outlined in the statistical guidelines of the *Psychonomic Society* (2012), “it is important to address the issue of statistical power. . . . Studies with low statistical power produce inherently ambiguous results because they often fail to replicate.” Despite such pleas, however, the issue of power has largely been neglected in psychological research so far. A possible reason is that the most common statistical procedure to control  $\alpha$  and  $\beta$  (i.e., the Neyman-Pearson procedure) often requires sample sizes much larger than those typically employed (Erdfelder, Faul, & Buchner, 1996). To illustrate, a two-tailed two-sample *t* test requires a total sample size of  $N = 210$  to detect a mean difference of medium size (i.e., Cohen’s  $d = .50$ ) with error probabilities  $\alpha = \beta = .05$ . In contrast, the common overall sample size for the same test is only about  $N = 60$  in prototypical journal publications, resulting in power values slightly lower than  $1 - \beta = .50$  (Cohen, 1962; Sedlmeier & Gigerenzer, 1989).

To avoid costly hypothesis tests, researchers may be tempted to apply NHST to small, underpowered samples first, followed by recursive increases in sample size until a significant test result is observed. This misleading use of NHST is known as data peeking, a questionable research practice that boosts chances of gaining a significant outcome at the cost of error probability control (Simmons, Nelson, & Simonsohn, 2011). In this article, we promote a proper alternative statistical method that was developed more than 70 years ago: sequential analysis (Wald, 1947). Unlike data peeking, with its associated risk of inflating Type I errors, sequential hypothesis tests have been designed specifically to control error probabilities while also allowing for smaller sample sizes than the Neyman-Pearson approach (Lakens, 2014). As computational tools have improved substantially over the past decades, these sequential tests are nowadays easily implemented and combined with standard statistical software. We empirically demonstrate the beneficial properties of one particular sequential test, namely, the sequential probability ratio test (SPRT; Wald, 1947). Moreover, we show that on top of controlling for decision error probabilities,

this test is more efficient than both the Neyman-Pearson approach and other common sequential designs. Importantly, we also assess the robustness of the proposed sequential test against violations of its assumptions.

The key feature of sequential tests, as opposed to standard test procedures, is that the sample size  $N$  is not determined a priori but a random variable that depends on the sequence of observations. Thereby, sequential methods may substantially reduce the sample size required to make a decision whenever the available data clearly support one hypothesis over the other. At the same time, they allow for explicit control of decision error probabilities. Thus, sequential statistical methods form an attractive alternative to standard test procedures. Despite their desirable properties and potential benefits to the field of psychological science, however, sequential methods have largely been ignored in experimental research so far (Botella, Ximénez, Revuelta, & Suero, 2006; Lakens, 2014; Lang, 2017).

One helpful step in this direction was recently taken by Schönbrodt, Wagenmakers, Zehetleitner, and Perugini (2017). These authors proposed a sequential method based on Bayesian inference, referred to as *sequential Bayes factors* (SBFs). By means of simulation, they demonstrated the properties of SBFs in the context of testing hypotheses about mean differences of two independent groups (two-sample *t* test). Specifically, they simulated populations with a specific mean difference  $\delta$  and examined the simulation estimate of the expected sample size and the relative frequencies of Type I and Type II errors of SBFs for different prior specifications and stopping criteria. Based on their simulations, they compared the SBF design with two other designs: the standard fixed-sample Neyman-Pearson *t* test (misleadingly referred to as the *null hypothesis significance test with power analysis* [NHST-PA]) and the group sequential (GS) design (Lang, 2017; Proschan, Lan, & Wittes, 2006).

In the GS design, the data are analyzed at predefined stages during the sampling process. If, in any stage, the test statistic exceeds a critical value, sampling is terminated. These critical values, in turn, are calculated based on linear spending functions of  $\alpha$  and  $\beta$  such that the overall error rates of the procedure can be controlled. Thus, while reducing the average sample size required for a statistical decision, the GS design does not compromise predefined error probabilities (Lakens, 2014). Nevertheless, Schönbrodt et al. (2017) showed that SBFs need smaller samples on average than both the Neyman-Pearson and the GS design in order to achieve the same error probabilities. Thus, they concluded that “SBF can answer the question about the presence or absence of an effect with better quality . . . and/or higher efficiency . . . than the classical NHST-PA approach or typical frequentist sequential designs” (p. 335).

We appreciate the contribution of Schönbrodt et al. (2017) in raising awareness for the advantages of sequential designs and thoroughly assessing the long-run properties of SBFs compared with the Neyman-Pearson and GS designs. However, their comparison did not include the arguably most efficient sequential design: the SPRT (Wald, 1947). We seek to close this gap and include the SPRT in the comparison. Moreover, as Schönbrodt et al. noted themselves, there is no means (and, in fact, no intention) in the standard SBF design to control statistical decision error probabilities explicitly. Herein, we will show that the SPRT not only allows for more efficient hypothesis tests about mean differ-

ences than SBFs and GS, it also exerts strict control of decision error probabilities.

In the following section, we briefly outline the basic concept of the SPRT, with particular focus on its application to the  $t$ -test scenario, elaborating on differences between the SPRT and other sequential designs. Next, we evaluate, by means of simulation, the properties of the SPRT with regard to empirical rates of incorrect decisions. Thereafter, we empirically compare SPRT, GS, and SBFs in terms of efficiency, that is, the expected sample size required to reach a decision. Subsequently, we explore the effects of violations of various assumptions underlying the test procedures. We then demonstrate the SPRT using an empirical example and discuss implications as well as limitations of our study and the SPRT. Finally, we provide recommendations on how to apply the proposed sequential  $t$  test in research practice.

### The Sequential Probability Ratio Test

Statistical tests usually assume samples of a fixed size  $N$ . Sequential statistical tests dispense with this requirement. Instead, the data are analyzed sequentially, and a rule is applied to make one of three possible decisions at any new step of the sampling process:

1. Accept  $\mathcal{H}_1$  and reject  $\mathcal{H}_0$ .
2. Accept  $\mathcal{H}_0$  and reject  $\mathcal{H}_1$ .
3. Continue sampling.

Whenever one of the first two decisions is made, the sampling process is terminated. In case of the third decision, another observation follows and the decision rule is applied again. This process is repeated until either one of the first two decisions is made. By implication, the sample size is not a fixed constant defined a priori but a random variable that depends on the sequence of observations.

To set up a sequential test, a decision rule needs to be defined. The choice of this rule determines the properties of the test, namely, the conditional probabilities of correct decisions and the so-called average sample number (ASN).<sup>1</sup> Assume that  $\mathcal{H}_0: \theta = \theta_0$  is tested against  $\mathcal{H}_1: \theta = \theta_1$ , where  $\theta$  denotes the true parameter (or parameter vector) in the underlying population. We shall impose on the test the following requirements (Wald, 1947):

$$P(\text{accept } \mathcal{H}_i | \theta_i) = \begin{cases} 1 - \alpha & (i = 0) \\ 1 - \beta & (i = 1) \end{cases}, \quad (2)$$

where  $P(\text{accept } \mathcal{H}_i | \theta_i)$  denotes the probability to correctly accept hypothesis  $\mathcal{H}_i$  when  $\theta_i$  is true. A sequential test is said to be of strength  $(\alpha, \beta)$  when it satisfies these requirements. For all tests of a given strength, a test is better if its ASN is smaller. Let  $E_{\theta_0}(N | S)$  denote the expected sample size  $N$  for a sequential test  $S$  when  $\theta$  is true. A test  $S'$  is better than an alternative test  $S$  of equal strength  $(\alpha, \beta)$  if  $E_{\theta_0}(N | S') < E_{\theta_0}(N | S)$  and  $E_{\theta_1}(N | S') \leq E_{\theta_1}(N | S)$ , or  $E_{\theta_0}(N | S') \leq E_{\theta_0}(N | S)$  and  $E_{\theta_1}(N | S') < E_{\theta_1}(N | S)$ . If there is a test  $S'$  such that for any alternative test  $S$  of equal strength  $E_{\theta_i}(N | S') \leq E_{\theta_i}(N | S)$ ,  $i = 0, 1$ , then  $S'$  is called an optimum test, because no other test of equal strength can exceed  $S'$  in terms of efficiency. For many applications, the choice of a decision rule to achieve an optimum test can be quite complex. However, for the special case of testing a simple null hypothesis against a simple alternative hypothesis, as in the given case, the SPRT has been proven to be optimal (Matthes, 1963; Wald & Wolfowitz, 1948).

Abraham Wald introduced the SPRT in the 1940s as one of the first formal theories of sequential test procedures. Let  $f(x | \theta_i)$  denote the probability (density) function for the observed data  $X$  given the population parameter specified in  $\mathcal{H}_i$ ,  $i = 0, 1$ . At any  $m$ th stage of the sampling process, compute a test statistic that conforms to the likelihood ratio, that is, the ratio of probability densities of the observed data  $X = x_1, \dots, x_m$  under  $\mathcal{H}_1$  versus  $\mathcal{H}_0$ , that is,

$$LR_m = \frac{f(x_1, \dots, x_m | \theta_1)}{f(x_1, \dots, x_m | \theta_0)}. \quad (3)$$

The likelihood ratio indicates how likely the observed data occur under one hypothesis vis-à-vis the other. It is thus a measure of relative evidence in the data for the specified hypotheses.<sup>2</sup> As a basis for statistical inference, it has desirable properties:

*Consistency:* The likelihood ratio is consistent, that is, if one of the specified hypotheses is in fact true, it will converge to either zero or  $\infty$  as the sample size increases toward infinity. Note that not all tests actually behave in this reasonable way. The  $p$  value in an NHST, for example, will not converge to 1 if the null hypothesis is true, which is why it is not suitable as a measure of evidence for the null (Rouder, Speckman, Sun, Morey, & Iverson, 2009).

*Independence from stopping rule:* Inference based on likelihood ratios is not affected by sampling plans and stopping rules (Etz, 2018). In NHST, statistical inference is based on the  $p$  value. This value is computed in reference to the sampling distribution of the test statistic under the null hypothesis and depends on the sample size. However, if the sample size is determined by what has been observed (*optional stopping*), the sampling distribution is likely to differ from the expected distribution under the assumption of a fixed sample size (Anscombe, 1954). Hence, its approximate properties (such as the  $p$  value) are unlikely to hold. Consequently, inference that is based on the assumption of a fixed sample size is affected by the stopping rule. The likelihood ratio, on the other hand, is independent of the researcher's intentions and stopping rule. Thus, it may be computed and interpreted sequentially (Etz, 2018).

Given these properties, Wald (1945, 1947) defined the following sequential test procedure based on the likelihood ratio:

1. Accept  $\mathcal{H}_1$  and reject  $\mathcal{H}_0$  when  $LR_m \geq A$ .
2. Accept  $\mathcal{H}_0$  and reject  $\mathcal{H}_1$  when  $LR_m \leq B$ .
3. Sample a new independent observation  $x_{m+1}$  when  $B < LR_m < A$ .

Wald (1947) showed that this sequential procedure terminates with probability 1 after a finite number of observations with either Decision 1 or 2. This implies that  $A \leq P(\text{accept } \mathcal{H}_1 | \theta_1) / P(\text{accept } \mathcal{H}_1 | \theta_0)$  and  $B \geq P(\text{accept } \mathcal{H}_0 | \theta_1) / P(\text{accept } \mathcal{H}_0 | \theta_0)$  (Wetherill, 1975). For practical purposes, these inequalities can be

<sup>1</sup> *Average sample number* denotes the average number of observations per sample, that is, the expected sample size at termination. Wald (1947) consistently used this expression; thus, we will maintain it as a technical term throughout the article.

<sup>2</sup> The term *likelihood* usually refers to the likelihood of a hypothesis,  $L(\mathcal{H})$ . This is proportional to the probability (density) of the data conditional on this hypothesis:  $L(\mathcal{H}) \propto f(x_1, \dots, x_n | \mathcal{H})$ . Thus, the likelihood ratio is usually expressed as a probability (density) ratio (Etz, 2018). Unlike Wald, however, we will maintain the term *likelihood ratio*.

replaced by equalities and, in accordance with the requirements given in Equation 2, the boundaries may simply be determined by  $A = (1 - \beta)/\alpha$  and  $B = \beta/(1-\alpha)$ . The resulting test will be approximately of strength  $(\alpha, \beta)$ : As the test statistic may exceed one of the boundaries at the point of termination rather than matching it exactly (a phenomenon called *overshooting*), the actual error probabilities of the sequential procedure will, in general, be lower than  $\alpha$  and  $\beta$ . Hence, strictly speaking, the SPRT is an approximate test, with  $\alpha$  and  $\beta$  serving as upper bounds to the error probabilities.

Importantly, this also holds true for interval hypotheses of the form  $\mathcal{H}_0: \theta \leq \theta_0$  versus  $\mathcal{H}_1: \theta \geq \theta_1$  ( $\theta_0 < \theta_1$ ) if all other parameters of the statistical model are known constants. Like the classical Neyman-Pearson test, an SPRT based on the simple hypotheses  $\theta = \theta_0$  versus  $\theta = \theta_1$  will have its maximum error probabilities  $\alpha$  and  $\beta$  if the true  $\theta$  equals  $\theta_0$  and  $\theta_1$ , respectively. For any other true value  $\theta$  in line with  $\mathcal{H}_0: \theta \leq \theta_0$  or  $\mathcal{H}_1: \theta \geq \theta_1$  ( $\theta_0 < \theta_1$ ), the respective error probabilities will be lower (Wald, 1947). Hence, just like Neyman-Pearson tests, SPRTs allow for the specification of upper-bound error probabilities even if there is no expectation of the exact value of the parameter of interest, as long as a minimum (maximum) value can be defined and all other parameters are constants.

### Sequential $t$ Tests

Despite the generality of the SPRT, a test procedure designed for decisions between simple hypotheses will not be appropriate for many applications (Wetherill, 1975). To see this, note that a hypothesis  $\mu = \mu_0$  on the mean of a normally distributed random variable would only be simple if the variance  $\sigma^2$  was either known or also specified by the hypothesis. If at least one of the parameters of the underlying statistical model is unknown, the decision becomes one between complex composite hypotheses to which the SPRT defined by Equations 3 and 4 does not apply. To adapt the SPRT to such hypotheses, Wald (1947) suggested the use of weight functions to integrate out the unknown parameters from the statistical model. However, the construction of suitable weight functions is not trivial. What is more, there is no general method such that the resulting SPRT satisfies the requirements concerning error probabilities and efficiency. In fact, the mathematical complexity of setting up suitable test statistics for composite hypotheses might in part be responsible for the widespread neglect of sequential methods in behavioral research (Botella et al., 2006).

Another way to cope with the problem of unknown parameters is to replace the sequence of observations in the likelihood ratio ( $LR_m$ ) by a transformed sequence that no longer depends on the unknown parameters (Armitage, 1947). For the one-sample test on the mean of a normal distribution with unknown variance, Barnard (1949) showed that composite hypotheses about  $\mathbf{X}$  can be reduced to simple hypotheses about the well-known  $t$  statistic computed from  $\mathbf{X}$ . Specifically, the sample observations  $x_1, \dots, x_m$  at stage  $m$  are simply replaced by the corresponding  $t$  statistics  $t_2, \dots, t_m$  based on these data ( $m \geq 2$ ), whose distributions do not depend on the unknown variance. Rushton (1950), building on previous work by Cox (1952), showed that an SPRT analogue of the one-sample  $t$  test can be performed by simply considering the ratio of probability densities for the most recent  $t_m$  statistic under  $\mathcal{H}_1$  and  $\mathcal{H}_0$  at any  $m$ th stage, because

$$LR_m = \frac{f(t_2, \dots, t_m | \mathcal{H}_1)}{f(t_2, \dots, t_m | \mathcal{H}_0)} = \frac{f(t_m | df_m, \Delta_1) \cdot f(t_2, \dots, t_{m-1} | t_m)}{f(t_m | df_m, \Delta_0) \cdot f(t_2, \dots, t_{m-1} | t_m)} = \frac{f(t_m | df_m, \Delta_1)}{f(t_m | df_m, \Delta_0)} \tag{5}$$

In Equation 5,  $df_m$  denotes the degrees of freedom and  $\Delta_i$  denotes the noncentrality parameter of the  $t$  distribution corresponding to hypothesis  $\mathcal{H}_i$  at the  $m$ th stage. For two-sided tests,  $t_m$  can be substituted by  $t_m^2$ , and  $LR_m$  is thus expressed as the ratio of  $t^2$  density functions (Rushton, 1952).

For testing mean differences between two independent samples with unknown variance (two-sample  $t$  test), Hajnal (1961) introduced an SPRT based upon the same principle. Let  $\delta = (\mu_1 - \mu_2)/\sigma$  denote the true standardized difference of means of the populations underlying the two groups (i.e., Cohen's  $d$  in the population), with  $\sigma$  representing the common (but unknown) population standard deviation. Assume a two-sided test of the hypothesis  $\mathcal{H}_0: \delta = 0$  against  $\mathcal{H}_1: \delta = d, d \neq 0$ . For each step  $m$  of the sampling process, let  $n_1$  and  $n_2$  be the number of observations in Group 1 and Group 2, respectively, such that  $m = n_1 + n_2$ . If observations from both populations underlying the groups and at least two different observations from the same group have been sampled (such that the sample estimate of the standard error becomes larger than zero), we compute

$$t_m^2 = \left( \frac{\bar{X}_{1m} - \bar{X}_{2m}}{\hat{\sigma}_m \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right)^2, \tag{6}$$

with the group means  $\bar{X}_{1m}$  and  $\bar{X}_{2m}$  in step  $m$  and the pooled standard deviation

$$\hat{\sigma}_m = \sqrt{\frac{(n_1 - 1) \cdot s_{1m}^2 + (n_2 - 1) \cdot s_{2m}^2}{n_1 + n_2 - 2}}, \tag{7}$$

where  $s_{1m}^2$  and  $s_{2m}^2$  denote the group variances estimated from the observed sample data available in step  $m$ .

The likelihood ratio is then derived as the ratio of the noncentral to the central probability density of  $t_m^2$ ,

$$LR_m = \frac{f(t_m^2 | df_m, \Delta_m)}{f(t_m^2 | df_m)}, \tag{8}$$

with  $df_m = n_1 + n_2 - 2$  and noncentrality parameter

$$\Delta_m = d \cdot \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}}. \tag{9}$$

Because  $t^2(df) = F(1, df)$ , the ratio in Equation 8 can be expressed as the ratio of a noncentral to a central  $F$  density function,

$$LR_m = \frac{f(F_m | d_1 = 1, d_2 = df_m, \Delta_m^2)}{f(F_m | d_1 = 1, d_2 = df_m)}, \tag{10}$$

where  $F_m = t_m^2$  and  $d_1$  and  $d_2$  denote the degrees of freedom of the  $F$  distribution.

Both the  $t$  and the  $F$  density function are available in the standard R environment (R Core Team, 2017). Thus, an SPRT for

a one- or two-sample  $t$  test can be conducted easily with R by iteratively computing the ratios given in Equation 5 for a one-sided test, and Equation 10 for a two-sided test, for each stage  $m$  of the sequential sampling process. A workable R script to apply the SPRTs described in this section can be downloaded from the Open Science Framework (<https://osf.io/4zub2/>).

Hajnal (1961) proved that a sequential procedure based on Equation 10 with the boundary values  $A = (1 - \beta)/\alpha$  and  $B = \beta/(1 - \alpha)$  results in a valid SPRT as described in the previous paragraph. Thus, the two-sample SPRT  $t$  test (henceforth referred to as Hajnal's  $t$  test) constitutes an easy to implement alternative to Neyman-Pearson tests as well as to SBFs and GS for the scenario addressed in Schönbrodt et al. (2017). In addition, it also provides full control of the error probabilities  $\alpha$  and  $\beta$ . However, the formal proof of the optimum property of the SPRT as well as analytical methods to determine the ASN of the procedure only apply to simple hypotheses and independent observations (Cox, 1952; Köllerström & Wetherill, 1979). Although Hajnal's  $t$  test transforms the composite hypothesis about  $X$  to a simple hypothesis about  $t$ , the sequence of  $t$  values is no longer composed of independent elements. Hence, neither the formal proof of the procedure's optimum character nor analytical solutions to determine the ASN hold for this test (Hajnal, 1961).

Therefore, it is of great practical as well as theoretical interest to empirically assess the properties of Hajnal's  $t$  test and examine (a) the degree to which the actual error rates approximate the upper bounds  $\alpha$  and  $\beta$ , (b) the expected sample size and relative efficiency compared with Schönbrodt et al.'s (2017) SBFs and the GS design, and (c) the robustness of these results when basic assumptions are violated. In the following section, we elaborate on the differences between the SPRT and the two alternative sequential test procedures addressed in this article.

## Two Alternative Sequential Designs: GS and SBFs

As outlined before, the GS is based on a priori planned stops during the sampling process. These stops include a number of interim tests and a final test, for which the sample size ( $N_{max}$ ) may be defined by a power analysis. For example, a researcher might decide to perform three interim tests after  $n = 25, 50,$  and  $75$  observations, say, before performing a final test at  $N_{max} = 100$  observations. Based on the overall error rates of the procedure,  $\alpha$  and  $\beta$ , critical values for the fixed-sample test statistic are calculated for each stop using linear spending functions (Lakens, 2014). The researcher will then sample 25 observations and compare the test statistic at this point with the critical values for the first analysis. If there is strong evidence in the data and the statistic exceeds a critical value, sampling is terminated and the respective hypothesis is accepted. Otherwise, the researcher has to continue sampling until the next stop is reached. This continues until  $N_{max}$ , where the test will finally accept one of the hypotheses.

Due to the interim analyses and the resulting possibility to terminate early, the GS requires on average fewer observations than Neyman-Pearson tests with the same error probabilities (Lang, 2017; Schönbrodt et al., 2017). Importantly, it allows for explicit control of these probabilities and the specification of a maximum number of observations required. As the interim analyses have to be planned a priori, however, the GS is less flexible than the SPRT or SBFs. Whereas the latter allow for termination

after possibly any single additional observation, a GS test can only terminate at one of the planned stops. Hence, although it has the advantage of a definite upper limit to the required sample size, it can be expected that the GS is, on average, less efficient than SPRT and SBFs (see Schönbrodt et al., 2017).

The test statistic of the SBF design is the Bayes factor (Jeffreys, 1935, 1961; Wrinch & Jeffreys, 1921). Like the SPRT test statistic, the Bayes factor is a likelihood ratio. Thus, it is a measure of relative evidence in the data for the specified hypotheses (Kass & Raftery, 1995):

$$BF_{10} = \frac{f(x_1, \dots, x_n | \mathcal{H}_1)}{f(x_1, \dots, x_n | \mathcal{H}_0)}. \quad (11)$$

Importantly, the likelihoods specified in this ratio are *marginal* likelihoods, that is, the probability density of data under hypothesis  $\mathcal{H}$  is given by

$$f(x_1, \dots, x_n | \mathcal{H}) = \int_{\Theta_{\mathcal{H}}} f_{\mathcal{H}}(x_1, \dots, x_n | \theta) p_{\mathcal{H}}(\theta) d\theta. \quad (12)$$

In Equation 12,  $\Theta_{\mathcal{H}}$  is the parameter space specified by hypothesis  $\mathcal{H}$ ,  $f_{\mathcal{H}}(x_1, \dots, x_n | \theta)$  is the probability density of the data given a certain point  $\theta$  in  $\Theta_{\mathcal{H}}$ , and  $p_{\mathcal{H}}(\theta)$  is the prior distribution of the parameters  $\theta$  under hypothesis  $\mathcal{H}$ . Thus, the likelihood is integrated over all possible values in the parameter space defined by the hypothesis, weighted according to the respective prior functions. In other words, the likelihood ratio in the Bayes factor is a weighted average of likelihood ratios for all possible parameter values (Morey & Rouder, 2011; Rouder et al., 2009).

In their simulation of the SBFs, Schönbrodt et al. (2017) used the default prior specifications as proposed by Jeffreys (1961) and Zellner and Siow (1980), which were further developed by Rouder et al. (2009) for the standard Bayesian  $t$  test. Specifically, prior distributions are defined for the unknown population variance, the grand mean, and the effect size, that is, the true standardized mean difference  $\delta$ . The likelihood under the null hypothesis is the likelihood for the constant  $\delta = 0$ , as in the SPRT. Under the alternative hypothesis, however, the specified prior for the effect size is not a constant but a Cauchy distribution whose shape is defined by a scale parameter  $r$ . Consequently, the Bayes factor tests the point hypothesis  $\mathcal{H}_0: \delta = 0$  against the alternative  $\mathcal{H}_1: \delta \sim \text{Cauchy}(r)$ .<sup>3</sup> With increasing scale parameter, the Cauchy distribution gets flatter, thus putting more weight on larger effect sizes. The default values suggested in the BayesFactor package in R for the test of a small, medium, or large effect are  $r = \sqrt{2}/2, 1,$  or  $\sqrt{2}$ , respectively (Morey & Rouder, 2015).

The likelihood ratios employed in the SPRT and SBFs are closely related. Unlike in the Bayesian  $t$  test, however, the alternative hypothesis in Hajnal's  $t$  test specifies a constant  $d$  rather than a distribution. Figure 1 illustrates how the probability density of observed data under the alternative hypothesis changes when marginalizing across an effect size prior distribution: Assume a hypothesis test on the mean difference of two normally distributed variables with some common, known variance. The probability density of an observed mean difference  $\hat{\delta}$  under the null hypothesis

<sup>3</sup> Note, however, that both hypotheses are composite hypotheses because of the unknown within-groups variance for which a common standard prior is assumed, known as *Jeffreys prior*, and the unknown grand mean, for which a uniform prior is specified (Rouder et al., 2009).

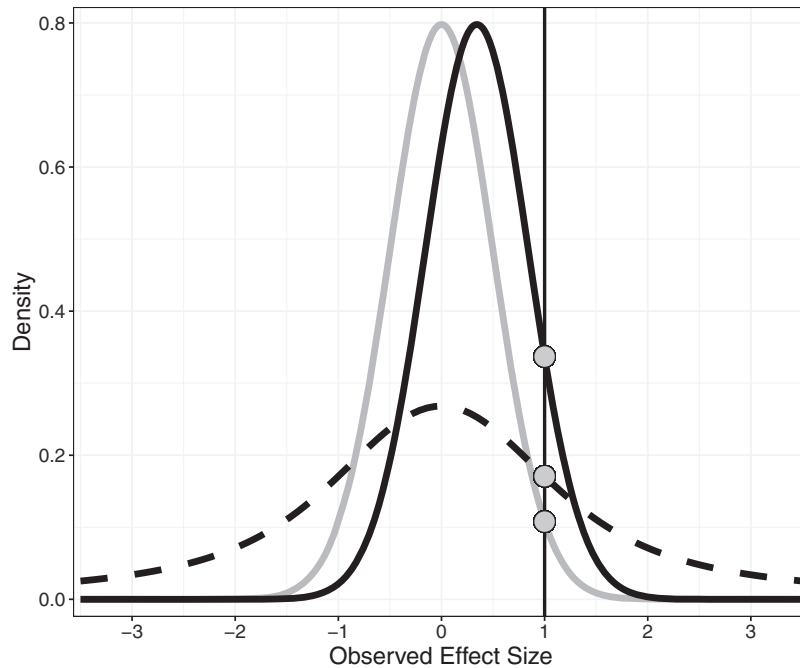


Figure 1. Effects of marginalizing across an effect-size prior, assuming some known variance. The gray line denotes the probability density of an observed effect size under the null hypothesis  $\delta = 0$ . The solid black line denotes the probability density of observed data under the alternative hypothesis  $\delta = d, d = 0.8$ . The dashed line denotes the density function when marginalized corresponding to the hypothesis  $\delta \sim \text{Cauchy}(1)$ . Gray dots denote the densities under either hypothesis for an observed effect of size  $\hat{\delta} = 1$ .

$\mathcal{H}_0: \delta = 0$  is given by the gray curve. Let the alternative hypothesis be  $\mathcal{H}_1: \delta = d, d = 0.8$ . Then, the solid black line denotes the respective probability density of an observed mean difference under this hypothesis. Now assume a sample effect size of  $\hat{\delta} = 1$  is observed. A likelihood ratio is simply the ratio of densities at the point of observed data (denoted by the gray dots in Figure 1). For the alternative hypothesis  $d = 0.8$ , this ratio is thus computed between the solid black and the gray curve at  $\hat{\delta} = 1$ . In the Bayes factor, however,  $f(\hat{\delta} | \mathcal{H}_1)$  is a weighted average of the probability densities under each possible  $\delta$  in  $\mathcal{H}_1: \delta \sim \text{Cauchy}(r)$ , with  $r = 1$  in this example. Consequently, the resulting probability density function (dashed curve) is less peaked than the density function based on the hypothesis  $d = 0.8$ . Thus, the ratio for the observed effect is larger under the latter than under the former hypothesis.

Generally speaking, a likelihood ratio based on point hypotheses will be more sensitive to data that are likely under the hypotheses. Consequently, if we assume that there either is no effect or a specific effect of size  $\delta = d$ , then the SPRT should be a more sensitive and more efficient test to discriminate between these two hypotheses than SBFs.

It should be noted, however, that this sensitivity comes at a cost: If the true effect differs greatly from what was expected ( $\hat{\delta} = 3$ , say), the likelihood ratio for the point alternative hypothesis will be less pronounced than for the diffuse hypothesis. As a consequence, in such a case, the SPRT is likely to be less efficient, whereas an SBF based on a diffuse prior will be more robust. A similar point was recently made by Stefan, Gronau, Schönbrodt, and Wagenmakers (2019). According to these authors, an SBF based on an

informative prior is more efficient (or less error-prone) when the true effect lies within the prior’s highest density region. At the same time, however, the informative prior might be at a disadvantage if the true effect greatly deviates from this region. In other words, there is a general trade-off between peak efficiency when the true effect matches the expectation, and robustness when it does not. Conceptually, the effect size specified in the SPRT is the most extreme case of an informed prior. Hence, Stefan et al.’s conclusions also apply to the SPRT.

### Statistical Error Rates of SPRT, GS, and SBFs

As outlined above, in practical applications, the SPRT will be an approximate test procedure, where  $\alpha$  and  $\beta$  serve as upper bounds to the actual error rates (Wald, 1947). Thus, we empirically examined the properties of Hajnal’s  $t$  test by means of simulations, focusing on the empirical rates of wrong decisions in relation to the specified upper bounds  $\alpha$  and  $\beta$ . Additionally, we simulated a GS test and SBFs with default Cauchy priors to assess their error rates under the same population scenarios.

Note, however, that whereas Hajnal’s  $t$  test and the GS design are based on the assumption of a fixed underlying effect and the same nominal error rates, the default priors in the SBFs make quite different assumptions. In a Cauchy distribution with scale parameter  $r$ , 50% of the area under the curve lie in the interval  $[-r, r]$ . Thus, the default scale parameters used by Schönbrodt et al. (2017),  $r = \sqrt{2}/2, 1$ , and  $\sqrt{2}$ , correspond to expected median absolute effect sizes of  $\delta = 0.7, 1$ , and  $1.4$ , respectively. The absolute effect sizes corresponding to a small, medium, or large

effect in Hajnal's  $t$  test as well as GS and Neyman-Pearson tests, in contrast, are  $\delta = 0.2, 0.5,$  and  $0.8,$  respectively (Cohen, 1988). Thus, our results—like those of Schönbrodt et al.—should not be generalized to other SBF designs with different prior distributions (e.g., informative priors; Stefan et al., 2019) or other population scenarios (e.g., random effects).

### Settings of the Simulation

We drew random samples from two normal distributions with common variance  $\sigma^2 = 1$  and means  $\mu_1 = \delta$  ( $\delta = 0, 0.2, 0.4, 0.5, 0.6, 0.8, 1, 1.2$ ), and  $\mu_2 = 0$ . Starting at  $n_1 = n_2 = 2$ , we applied Hajnal's  $t$  test to the sample data. The sample of each group was then increased by +1 until the  $LR$  exceeded one of the boundary values  $A = (1 - \beta)/\alpha$  or  $B = \beta/(1 - \alpha)$ . In addition to the true effect size  $\delta$ , the settings of the test procedure were varied in terms of expected effect size  $d$  according to  $\mathcal{H}_1$  and typical values of the nominal error probabilities  $\alpha$  and  $\beta$ , that is,  $\alpha = .01$  versus  $.05$ , and  $\beta = .05$  versus  $.10$ . For each combination of true effect size  $\delta$ , expected effect size  $d$ ,  $\alpha$ , and  $\beta$ , 10,000 replications were simulated.

In a second step, we simulated a GS with four looks (three interim analyses and one final test) for the same population scenarios and nominal error rates. Sample sizes for each step and the respective critical values were calculated with the `gsDesign` package in R (K. Anderson, 2014).

Third, we replicated Schönbrodt et al.'s (2017) simulation of the SBFs: Random samples from two normally distributed populations with true mean difference  $\delta$  were drawn, and the Bayes factor was computed during the sampling process until a threshold value was reached. The scale parameter of the Cauchy prior in the Bayes factor was systematically varied using the default values specified in the `BayesFactor` package, that is,  $r = \sqrt{2}/2, 1,$  or  $\sqrt{2}$ , respectively (Morey & Rouder, 2015). The threshold values for the sequential procedure were set to a critical Bayes factor between 3 and 30 in steps of 1. As in the previous simulation of Hajnal's  $t$  test, each simulated trajectory started with an initial sample size of  $n_1 = n_2 = 2$  that was gradually increased in equal steps for both groups until a decision threshold was reached.<sup>4</sup>

### Results

Columns 1 to 4 of Table 1 contain the percentages (and 95% confidence intervals [CIs]) of decision errors of Hajnal's  $t$  test as a function of the true effect  $\delta$ , the expected effect  $d$  under  $\mathcal{H}_1$ , and the specified error probabilities  $\alpha$  and  $\beta$ . In Columns 5 to 8, the same information is presented for the simulated GS with four looks. The remaining columns provide the result for the SBFs as a function of the true effect  $\delta$ , the scale parameter  $r$  of the Cauchy prior (representing the expected median absolute effect size under  $\mathcal{H}_1$ ), and the threshold value for the Bayes factor.

For the sake of brevity, we display only a limited range of effect sizes here, namely,  $\delta = 0, 0.2, 0.5,$  and  $0.8,$  as these represent the absence of an effect and the effect sizes commonly referred to as small, medium, and large (Cohen, 1988). In a similar vein, we report only a subset of SBF threshold values, namely, 5, 10, and 30. The full table of results as well as reproducible scripts and all data can be downloaded from <https://osf.io/4zub2/>. The ASN (as well as the 50th, 75th, and 95th quantile) for Hajnal's  $t$  test, GS,

and SBFs corresponding to the results displayed in Table 1 may be obtained from the Appendix (Table A1).

The first three rows of Table 1 depict the observed percentages of incorrect decisions for the true population scenario  $\delta = 0$  (i.e., empirical Type I error rates). Obviously, Hajnal's  $t$  test provides excellent  $\alpha$  error control. The empirical rates closely approximate the nominal probabilities (.01, .05). In fact, as can be inferred from the 95% Clopper-Pearson exact CIs (Clopper & Pearson, 1934), 67% of the observed Type I error rates are significantly lower than the specified  $\alpha$ . Thus, as expected, Hajnal's  $t$  test approximates nominal error rates nicely, with the specified  $\alpha$  serving as an upper bound.

We observe a similar result for the GS: The empirical error rates nicely approximate the nominal  $\alpha$ . In some cases, the estimate is slightly above the nominal level, but this is likely caused by sampling error. Hence, with respect to Type I error control, the GS and the SPRT procedures are comparable and perform well.

In contrast to the SPRT and GS test procedures, the observed  $\alpha$  rates of the SBFs vary as a function of the Bayes factor threshold value and the scale parameter  $r$ . For a low threshold, the probabilities of falsely rejecting a true null hypothesis are much larger than what researchers typically aim at. Although these error rates decrease for higher thresholds (e.g., about .06 for a Bayes factor of 10), there is no means in the standard SBF design to control the  $\alpha$  probability a priori.

The remaining rows of Table 1 correspond to true population scenarios with  $\delta > 0$ . Here, the percentages represent observed rates of accepting a false null hypothesis (Type II error). The probability of committing such an error is commonly referred to as  $\beta$ ; however, this definition is somewhat vague. More precisely,  $\beta$  is the probability to accept a false null hypothesis if the specified alternative hypothesis  $\delta = d$  is in fact true (see Equation 2). As the results in Table 1 demonstrate, Hajnal's  $t$  test provides excellent control of the error probability in this situation: The empirical rates nicely approximate but never exceed the specified  $\beta$  (.05, .1). In fact, the actual error rates are significantly smaller than the nominal  $\beta$  in 92% of the cases. Thus, as expected,  $\beta$  denotes an upper bound of the test procedure's probability to accept a false null hypothesis when the alternative is correctly specified.

Notably, this result also holds when the true effect does not match the expected effect but is in fact larger. As Table 1 shows, the probability that Hajnal's  $t$  test incorrectly accepts a false null hypothesis converges to zero when  $\delta > d$ . It is a popular critique by proponents of the Bayesian approach that a precise prediction of the effect size is not possible (e.g., Schönbrodt et al., 2017). Even in this case, however, a test can be defined with  $\beta$  as an upper bound to the Type II error probability (Wald, 1947). By specifying a minimum relevant effect  $d_{min}$  and setting up Hajnal's  $t$  test for the simple hypothesis  $\mathcal{H}_0: \delta = 0$  against  $\mathcal{H}_1: \delta = d_{min}$ , the probability of incorrectly accepting a false null hypothesis will never exceed  $\beta$  if  $\delta \geq d_{min}$ . Of course, if the true effect is notably smaller than  $d_{min}$  then the probability of accepting  $\mathcal{H}_0$  will exceed  $\beta$ . However, if  $d_{min}$  is specified based on which effect sizes are practically relevant, one would actually prefer the test to maintain

<sup>4</sup> To find an acceptable compromise between computational efficiency and accuracy in the simulations of Hajnal's test and SBFs, the samples were increased by +1 until  $n_1 = n_2 = 10,000$  and by +50 afterward.

**Table 1**  
**Percentages (and 95% CI) of Type 1 and Type 2 Decision Errors Committed by Hajnal's t Test, Group Sequential Test, and Sequential Bayes Factors**

<i>d</i>	<i>r</i>	Hajnal's <i>t</i> test						Group sequential test						Sequential Bayes factors		
		$\alpha = 1\%$		$\alpha = 5\%$		$\alpha = 10\%$		$\alpha = 1\%$		$\alpha = 5\%$		$\alpha = 10\%$		BF = 5	BF = 10	BF = 30
		$\beta = 5\%$	$\beta = 10\%$	$\beta = 5\%$	$\beta = 10\%$	$\beta = 5\%$	$\beta = 10\%$	$\beta = 5\%$	$\beta = 10\%$	$\beta = 5\%$	$\beta = 10\%$	$\beta = 5\%$	$\beta = 10\%$			
$\delta = 0$ (% Type 1 error)																
.2	$\sqrt{2}/2$	1.0 [0.9, 1.3]	.9 [7.1, 1.1]	4.5 [4.1, 4.9]	4.8 [4.4, 5.3]	.8 [7.1, 1.0]	1.0 [0.8, 1.2]	5.1 [4.6, 5.5]	4.7 [4.3, 5.1]	11.6 [11.0, 12.3]	6.3 [5.9, 6.8]	2.3 [2.0, 2.6]				
.5	1	.8 [7.1, 1.0]	.7 [5.9, .9]	4.5 [4.1, 5.0]	3.8 [3.4, 4.2]	.8 [7.1, 1.0]	.9 [7.1, 1.1]	5.3 [4.9, 5.8]	4.9 [4.5, 5.3]	10.3 [9.7, 10.9]	5.8 [5.3, 6.3]	2.0 [1.8, 2.3]				
.8	$\sqrt{2}$	.8 [7.1, 1.0]	.8 [6.1, 1.0]	4.0 [3.6, 4.4]	3.8 [3.5, 4.2]	1.2 [1.0, 1.5]	1.3 [1.1, 1.5]	5.4 [5.0, 5.9]	5.6 [5.1, 6.0]	10.5 [9.9, 11.1]	5.8 [5.4, 6.3]	2.1 [1.8, 2.4]				
$\delta = .2$ (% Type 2 error)																
.2	$\sqrt{2}/2$	4.7 [4.3, 5.1]	9.4 [8.8, 10.0]	4.4 [4.0, 4.8]	9.1 [8.5, 9.6]	4.4 [4.0, 4.8]	9.7 [9.1, 10.3]	5.1 [4.7, 5.6]	9.1 [8.6, 9.7]	49.0 [48.0, 50.0]	6.4 [5.9, 6.9]	.0 [0.0, .0]				
.5	1	81.0 [80.2, 81.7]	84.8 [84.1, 85.5]	73.1 [72.2, 73.9]	74.6 [73.7, 75.5]	80.5 [79.7, 81.3]	84.2 [83.4, 84.9]	68.2 [67.3, 69.2]	73.6 [72.8, 74.5]	67.0 [66.0, 67.9]	23.5 [22.7, 24.4]	.0 [0.0, .0]				
.8	$\sqrt{2}$	95.2 [94.7, 95.6]	95.3 [94.9, 95.7]	88.1 [87.5, 88.8]	89.2 [88.5, 89.8]	92.7 [92.1, 93.2]	93.8 [93.4, 94.3]	85.2 [84.5, 85.9]	85.6 [84.9, 86.3]	78.6 [77.8, 79.4]	49.0 [48.1, 50.0]	.1 [0.0, .2]				
$\delta = .5$ (% Type 2 error)																
.2	$\sqrt{2}/2$	0 [0.0, .0]	0 [0.0, .0]	0 [0.0, .0]	0 [0.0, .0]	0 [0.0, .0]	0 [0.0, .0]	0 [0.0, .0]	0 [0.0, .0]	1.7 [1.4, 1.9]	0 [0.0, .0]	.0 [0.0, .0]				
.5	1	4.2 [3.8, 4.6]	8.8 [8.2, 9.4]	4.5 [4.1, 4.9]	8.7 [8.2, 9.3]	4.2 [3.8, 4.6]	8.6 [8.1, 9.2]	4.7 [4.2, 5.1]	8.5 [8.0, 9.1]	11.7 [11.1, 12.3]	.0 [0.0, .1]	.0 [0.0, .0]				
.8	$\sqrt{2}$	39.8 [38.8, 40.8]	48.8 [47.8, 49.8]	35.4 [34.4, 36.3]	43.6 [42.6, 44.6]	43.5 [42.5, 44.5]	52.9 [51.9, 53.9]	35.4 [34.5, 36.4]	44.7 [43.7, 45.6]	31.8 [30.9, 32.7]	1.3 [1.1, 1.5]	.0 [0.0, .0]				
$\delta = .8$ (% Type 2 error)																
.2	$\sqrt{2}/2$	0 [0.0, .0]	0 [0.0, .0]	0 [0.0, .0]	0 [0.0, .0]	0 [0.0, .0]	0 [0.0, .0]	0 [0.0, .0]	0 [0.0, .0]	0 [0.0, .0]	0 [0.0, .0]	.0 [0.0, .0]				
.5	1	0 [0.0, .1]	1 [0.0, .2]	0 [0.0, .1]	2 [1.1, .3]	0 [0.0, .0]	0 [0.0, .1]	0 [0.0, .1]	.1 [0.0, .2]	6 [4.8, .8]	0 [0.0, .0]	.0 [0.0, .0]				
.8	$\sqrt{2}$	3.9 [3.5, 4.3]	7.7 [7.2, 8.2]	4.1 [3.7, 4.5]	8.2 [7.7, 8.8]	4.5 [4.1, 4.9]	8.0 [7.4, 8.5]	4.1 [3.7, 4.5]	8.7 [8.2, 9.3]	5.9 [5.4, 6.4]	.0 [0.0, .1]	.0 [0.0, .0]				

*Note.* The *d* and *r* metrics indicate fundamentally different effect-size expectations, although they are often assigned the same verbal labels for “small” ( $d = .2$ ,  $r = \sqrt{2}/2$ ), “medium” ( $d = .5$ ,  $r = 1$ ), and “large” effects ( $d = .8$ ,  $r = \sqrt{2}$ ). The group sequential test comprised three interim and one final test. *d* = expected effect size according to  $\mathcal{H}_1$  in Hajnal's *t* test and group sequential test (Cohen's *d*); *r* = scale parameter of Cauchy prior (= expected median absolute effect size) according to  $\mathcal{H}_1$  in sequential Bayes factors;  $\delta$  = true population effect size (Cohen's *d* in population); BF = threshold Bayes factor.



$\mathcal{H}_0$  if the true effect falls under this lower bound. Thus, the results demonstrate that Hajnal's  $t$  test provides reliable, conservative control of the probabilities to commit a decision error.

It should be noted at this point, however, that a conservative specification of the effect results not only in conservative error rates but also in a less efficient test: In the same way as error rates decrease when the true effect is larger than expected, the ASN increases (Table A1). This is not surprising, as this reflects the trade-off between efficiency and robustness (Stefan et al., 2019). Importantly, this is also true for the GS and the classical Neyman-Pearson test. As the effect-size assumption is the same for all three designs, a conservative estimate will increase the required sample size for all of them. As Table A1 shows, however, Hajnal's  $t$  test is still more efficient in these cases. For example, if a small effect is expected ( $d = 0.2$ ) in case of a medium true effect ( $\delta = 0.5$ ), Hajnal's  $t$  test with  $\alpha = \beta = .05$  requires, on average, 194 observations. The GS test with the same parameters requires, on average, 378 observations. A classical  $t$  test with the same assumptions would even require 1,302 observations. Hence, Hajnal's  $t$  test is more efficient not only when the correct effect size is expected but also when the tests specify a conservative assumption.

As in case of Type I errors, there is no explicit control of Type II errors in the SBF design. For a threshold Bayes factor of 5, the empirical error rates exceed typical error rates by far (see also Schönbrodt et al., 2017). A more reasonable threshold of 10 yields excellent error probabilities for medium to large effects but not for small effects ( $\delta = 0.2$ ). If a higher threshold is chosen ( $\text{BF} = 30$ ), the procedure will basically commit no decision errors, even when the true effect is small. However, this powerful procedure comes at the cost of efficiency: In the context of small to medium effect sizes, the expected sample sizes required to reach the decision threshold can become extremely large (see Table A1). For example, for a true effect of size  $\delta = 0.2$ , an SBF assuming a Cauchy prior with scale  $r = \sqrt{2}/2$  requires, on average, 1,120 observations to reach a threshold of 30. To summarize, the results indicate that the SBF design—if combined with thresholds representing moderate ( $\text{BF} = 5$ ) or strong evidence ( $\text{BF} = 10$ )—can be associated with high error probabilities and lacks a proper means to control these explicitly.

### Relative Efficiency of SPRT, GS, and SBFs

For the test of simple hypotheses, the SPRT's properties can be derived analytically and its optimum character has been proven (Wald & Wolfowitz, 1948). When modified for the case of a composite hypothesis, however, analytical solutions no longer exist (Cox, 1952; Hajnal, 1961; Köllerström & Wetherill, 1979). Schönbrodt et al. (2017) demonstrated that the SBF design is more efficient for the two-sample  $t$  test scenario than the GS. However, this comparison did not include the SPRT, and sensitivity considerations concerning SBFs and SPRT strongly suggest that if SPRT's assumptions are met (which was the case in Schönbrodt et al.'s simulation design), it should be more efficient. To assess this in more detail, we empirically juxtaposed Hajnal's  $t$  test with SBFs and GS by means of simulation.

### Settings of the Simulations

A meaningful comparison of different test procedures' efficiencies requires all tests to satisfy the same error probabilities. To generate tests of the same strength ( $\alpha, \beta$ ), we repeated the simulation of Hajnal's  $t$  test with the same settings as in the previous simulation. This time, however, the stopping thresholds  $A$  and  $B$  were based on the corresponding error rates of the SBFs. For each condition, the test was based on the correctly specified effect-size assumption  $d = \delta$  and the empirical  $\alpha$  and  $\beta$  of the SBFs under the same condition.<sup>5</sup> In addition, we calculated the ASN for a GS test with four looks using the `gsDesign` package as well as required sample sizes for the corresponding Neyman-Pearson  $t$  test ( $N_{NP}$ ) with the same error probabilities. Thus, the four test procedures are of the same strength and can be compared directly in terms of efficiency.

Note that this simulation represents a favorable scenario for the SPRT, the GS, and the Neyman-Pearson test, as the true effect sizes perfectly match the effect-size assumptions. Hence, the results capture their peak efficiency. If the true effect does not match the expected effect or if different priors are used in the SBFs, the results are likely to differ. However, this simulation setting is necessary to keep the error rates constant across test procedures, which, in turn, is necessary for a meaningful comparison of efficiency.

### Results

The relative efficiency of Hajnal's  $t$  test, the GS, and the SBFs with default priors can be obtained from Figure 2. It is based on Figure 4 in Schönbrodt et al. (2017, p. 331), in which these authors presented their comparison of SBFs and GS. Figure 2 displays the relative reduction of the ASN of the three test designs compared with the corresponding Neyman-Pearson sample size (in %  $N_{NP}$ ). Not surprisingly, all three sequential designs are more efficient than the classical Neyman-Pearson  $t$  test. We also replicated the finding of Schönbrodt et al. that the SBF design (dashed line) is substantially more efficient than the GS test (dotted line), although the latter assumes the correct effect size. The mean relative reduction of expected sample size of the SBFs compared with the corresponding Neyman-Pearson test is 63%, whereas the mean relative reduction is 50% for the GS. However, as Figure 2 also reveals, Hajnal's  $t$  test is in fact even more efficient (solid line): On average, the ASN of Hajnal's  $t$  test is 67% smaller than  $N_{NP}$ . In almost all conditions, the observed ASN undercuts the corresponding statistics of the SBFs and the GS.

As can be seen, this difference between SPRT and SBFs is quite small for medium to large effect sizes, although consistent. Two mechanisms can explain this small difference:

1. As the true effect further departs from the null, the likelihood of an observation that is typical under the null hypothesis decreases quickly. Thus, the expected change in the likelihood ratio by adding a single observation increases correspondingly fast. Therefore,

<sup>5</sup> In conditions for which the SBFs did not exhibit wrong decisions,  $\beta$  was set to an arbitrarily small value of 1/50,000.

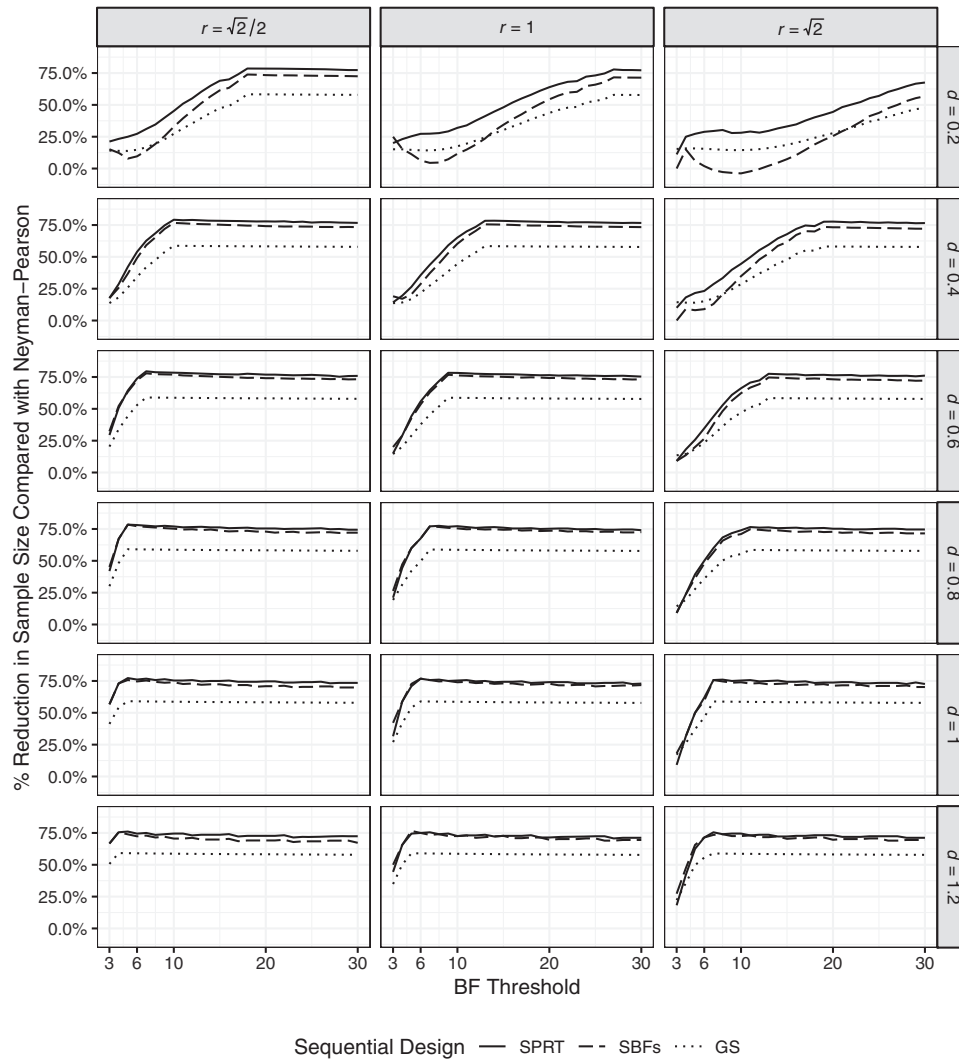


Figure 2. Relative efficiency of sequential probability ratio test (SPRT), group sequential test (GS), and sequential Bayes factors (SBFs). The y-axis denotes the reduction in expected sample size of SPRT (solid line), GS (dotted line) and SBFs (dashed line) compared with a Neyman-Pearson  $t$  test with the same error probabilities in  $\% N_{NP}$  for different true effect sizes as well as boundaries and prior specifications of the SBFs. Based on Schönbrodt et al. (2017, Figure 4).

both sequential procedures reach a decision on average after very few observations already (e.g., for  $\delta = 1.2$ , all ASNs are smaller than 20). Hence, the comparison is distorted in this context by a floor effect.

- The extent to which the likelihood ratio exceeds the stopping values  $A$  and  $B$  at the point of termination (*overshooting*) increases as a function of effect size (Wald, 1947). Thus, with increasing effect size, the actual error probabilities of the SPRT further depart from the specified  $\alpha$  and  $\beta$ , resulting in a more conservative and slightly less efficient test. Wald conjectured that this loss of efficiency was not of practical relevance. Nevertheless, it is important to keep in mind when interpreting the small difference in effi-

ciency between Hajnal's  $t$  test and SBFs in the context of large effect sizes.

For small to medium effect sizes, in contrast, the discrepancy in efficiency between Hajnal's  $t$  test and the SBFs is considerably stronger and can reach differences in ASNs of more than 300 observations. Because these are effect sizes that require very large  $N_{NP}$ s, efficiency is of particular interest in this context. Thus, as our results show, Hajnal's  $t$  test can be an efficient alternative not only with respect to the Neyman-Pearson procedure and the GS but also with respect to the default SBF design proposed by Schönbrodt et al. (2017). What is more, unlike the latter, Hajnal's  $t$  test additionally allows for the proper specification of upper bounds to decision error probabilities, as empirically illustrated in the previous section.

### Robustness of SPRT, GS, and SBFs

So far, we examined Hajnal’s  $t$  test under ideal conditions, that is, when the assumptions underlying the test procedure are met. This is necessary from a theoretical perspective in order to investigate the general properties of the test such as error probability control and efficiency, and also to compare the test procedure with other designs. However, from a practical point of view, it is also important to consider scenarios in which these assumptions are violated.

H. Lee and Fung (1980) already examined the robustness of Hajnal’s  $t$  test under conditions of non-normality and heteroscedasticity. Due to computational limitations at the time, however, their simulations were based on approximations to the likelihood ratio. In this section, we examine the performance of Hajnal’s  $t$  test as well as the GS and the default SBFs under conditions of (a) non-normality and (b) heteroscedasticity, as well as (c) random effects and (d) intentional misuse. For the sake of parsimony, we restricted the simulations to the nominal error rates  $\alpha = \beta = .05$  in Hajnal’s  $t$  test and the GS. For the SBFs, we chose a threshold value of  $BF = 10$  throughout the simulations. As this value reflects “strong evidence” from a Bayesian perspective (M. D. Lee & Wagenmakers, 2013), it is often used as a threshold in practical applications (e.g., Matzke et al., 2015; Schönbrodt et al., 2017; Wagenmakers et al., 2015). In each simulation, 10,000 replications per parameter combination were simulated. All scripts and data are again available from the Open Science Framework.

### Non-Normality

**Settings.** To investigate the test procedures’ performances against violations of the normality assumption, we repeated the first simulation for data generated from log-normal distributions and mixtures of two normal distributions. For the former case, we drew random data for two groups from a log-normal distribution corresponding to a standard normal on the log scale. To each observation in the first group,  $\delta\sigma'$  was added, where  $\delta$  denotes the

true standardized mean difference ( $\delta = 0, 0.2, 0.5, 0.8$ ) and  $\sigma'$  represents the standard deviation of the log-normal distribution.

To simulate the mixture case, we followed the procedure employed by H. Lee and Fung (1980) by generating random data from a mixture of two normal distributions given by

$$\gamma\mathcal{N}(\mu_1, \sigma_1) + (1 - \gamma)\mathcal{N}(\mu_2, \sigma_2), \tag{13}$$

where  $\gamma$  ( $\gamma = .9, .7, .5$ ) denotes the probability that an observation is drawn from  $\mathcal{N}(\mu_1, \sigma_1)$ . For the underlying distributions, we defined  $\mu_1 = 0, \sigma_1 = 1, \mu_2 = 2, \text{ and } \sigma_2 = 2$ . As in the log-normal case,  $\delta\sigma'$  was added to each observation in the first group, with  $\sigma'$  denoting the standard deviation of the mixture distribution.

**Results.** The empirical error rates of the three test procedures are displayed in Table 2. Expected sample sizes can be obtained from the Appendix (Table A2). For ease of comparison, the first two rows of Table 2 contain results from the first simulation for normally distributed data (see Table 1, Columns 3, 7, and 10).

In terms of error rates, the examined procedures are quite robust against violations of distributional assumptions. This is not surprising as it is in line with Lee and Fung’s results for Hajnal’s  $t$  test. Type I error rates in particular seem to be quite stable for all designs across all simulated scenarios, although Hajnal’s  $t$  test becomes slightly conservative with increasing effect-size assumption for log-normally distributed data. In the case of mixture distributions, there is a tendency that Type II error rates for all test procedures decrease with decreasing kurtosis. Interestingly, however, this is not accompanied by an increase in expected sample sizes. To summarize, all three sequential designs show robustness under conditions of non-normality both in terms of error rates and ASNs.

### Heteroscedasticity

**Settings.** To simulate the case of two populations with unequal variance and some standardized mean difference  $\delta$ , we drew random samples from two normal distributions with  $\mu_1 = 0, \sigma_1 = 1$  and

Table 2

*Percentages of Type 1 and Type 2 Decision Errors Committed by Hajnal’s  $t$  Test, Group Sequential Test, and Sequential Bayes Factors Under Conditions of Non-Normality*

Distribution	$\gamma$	$s$	$k$	Empirical error rates	$\delta = .2$			$\delta = .5$			$\delta = .8$		
					SPRT	GS	SBF	SPRT	GS	SBF	SPRT	GS	SBF
Normal		.0	3.0	$\alpha'$	4.5	5.1	6.3	4.5	5.3	5.8	4.0	5.4	5.8
				$\beta'$	4.4	5.1	6.4	4.5	4.7	.0	4.1	4.1	.0
Mixture	.9	.8	6.0	$\alpha'$	4.5	3.3	5.5	3.7	4.3	5.1	3.5	4.4	5.1
				$\beta'$	4.5	3.9	5.6	3.9	3.7	.0	4.1	4.2	.0
	.7	.9	4.4	$\alpha'$	4.6	2.5	6.2	4.2	2.5	4.6	3.7	3.1	4.5
				$\beta'$	2.9	2.0	4.2	2.9	1.8	.0	2.6	1.8	.0
	.5	.7	3.4	$\alpha'$	2.3	1.2	5.2	3.8	1.5	5.1	4.0	1.7	4.7
				$\beta'$	.8	.4	1.6	1.3	.4	.0	1.4	.3	.0
Log-normal	6.2	116.9	$\alpha'$	3.9	4.8	4.0	2.7	4.7	3.7	1.8	4.6	3.3	
			$\beta'$	4.6	5.1	6.2	4.1	4.5	.1	3.7	4.5	.0	

*Note.* The first two rows display results from the first simulation for normally distributed data (see Table 1, Columns 3, 7, and 10). Number of repetitions per parameter combination:  $k = 10,000$ .  $\gamma$  = mixture probability;  $s$  = skewness;  $k$  = kurtosis;  $\delta$  = true and expected effect size (Cohen’s  $d$  in population); SPRT = sequential probability ratio test (Hajnal’s  $t$  test) assuming  $d = \delta$  and  $\alpha = \beta = .05$ ; GS = group sequential design with four tests, assuming  $d = \delta$  and  $\alpha = \beta = .05$ ; SBF = sequential Bayes factor design with threshold 10, assuming  $r = \sqrt{2}/2, 1, \sqrt{2}$  when  $\delta = .2, .5, \text{ and } .8$ , respectively.

$$\mu_2 = \delta \cdot \sqrt{\frac{\sigma_2^2 + 1}{2}}, \tag{14}$$

with  $\sigma_2 = 1/4, 1, 4$  and  $\delta = 0, 0.2, 0.5, 0.8$ . In addition, we simulated two sampling schemes (H. Lee & Fung, 1980): (a) pairwise sampling from the two populations such that  $n_1/n_2 = 1$ , and (b) unbalanced sampling such that at each step for one observation in the first sample there were always three in the second sample, that is,  $n_1/n_2 = 1/3$ .

**Results.** The observed error rates for the three test procedures under the condition of heteroscedasticity are displayed in Table 3. The corresponding expected sample sizes can be obtained from the Appendix (Table A3). If the sample sizes are balanced, Hajnal's *t* test is basically unaffected by heteroscedasticity in the underlying populations. Although there seems to be a slight tendency that with increasing expected effect size, the Type I error rate increases as well, this is likely due to sampling error. In the same vein, expected sample sizes of Hajnal's *t* test are basically constant irrespective of the variance ratio as long as the group sample sizes are balanced. Thus, our simulations show that for a balanced sampling scheme, Hajnal's *t* test is robust against violations of homoscedasticity assumptions.

The GS seems to be quite robust as well (there is virtually no effect in terms of efficiency), although its empirical Type I error rates slightly exceed the nominal level. The SBF design is quite robust when there is an effect (Type II errors), but there is a noticeable increase in Type I error rates in the case of unequal variances. Thus, SBFs seem to be affected by heteroscedasticity to a certain extent even when group sample sizes are balanced.

If sample sizes are unbalanced, Hajnal's *t* test and the SBFs are affected in quite the same manner. If there is homoscedasticity, error rates do not change, whereas their efficiency is lowered: Expected sample sizes of both tests increase notably. In the case of heteroscedasticity, however, both tests show poor Type I error rates when the sample with larger variance is smaller. This increase in error rates is not surprising, as the pooled variance estimate will seriously underestimate the true variance if the sam-

ple with larger variance is notably smaller than the other sample. This, in turn, will result in too large *t* values and a high number of false-positive decisions. If the population with larger variance is overrepresented, on the other hand, both tests become more conservative and less efficient.

Interestingly, the GS is affected most seriously by an unbalanced design. Whereas Type I error rates are highly conservative when there is heteroscedasticity and the sample with small variance is larger, Type II error rates are inflated for all variance ratios. Hence, independent from heteroscedasticity, the GS design is strongly affected by unequal sample sizes.

**Random Effects**

**Settings.** In the previous simulations, a fixed effect size  $\delta$  was always assumed. This is a common assumption in psychology; however, it is also possible to assume that in certain cases, the true effect is not fixed but in fact random. Hajnal's *t* test, like the GS and the classical *t* test, specifies a fixed effect size. The default SBFs, on the other hand, are based on an effect-size prior distribution. Therefore, we investigated the performance of the three sequential designs when the true effect is in fact sampled from a distribution.

In this simulation, a population effect size  $\delta$  was randomly drawn from a normal distribution with  $\mu_\delta = 0.2, 0.5, 0.8$  and  $\sigma_\delta = 1$  in a first step. Subsequently, random data were drawn from two normal distributions with  $\mu_1 = \delta, \mu_2 = 0$  and common standard deviation  $\sigma = 1$ . In the case of  $\mathcal{H}_1$ , the expected effect size in Hajnal's *t* test and the GS was specified as  $d = \mu_\delta$ . In the SBFs, a Cauchy prior with  $r = \mu_\delta$  was specified. With this setting, the median expected absolute effect size in the SBFs always matches the true median effect size ( $\mu_\delta$ ) and, thus, in contrast to the fixed-effects simulations, this represents a favorable setting for the Bayesian test.

**Results.** The empirical error rates and ASNs can be obtained from Table 4. Not surprisingly, the error rates of Hajnal's *t* test and the GS are basically equivalent because they make the same

Table 3  
Percentages of Type 1 and Type 2 Decision Errors Committed by Hajnal's *t* Test, Group Sequential Test, and Sequential Bayes Factors Under Conditions of Heteroscedasticity

$N_1/N_2$	$\sigma_1/\sigma_2$	Empirical error rates	$\delta = .2$			$\delta = .5$			$\delta = .8$		
			SPRT	GS	SBF	SPRT	GS	SBF	SPRT	GS	SBF
1	1/4	$\alpha'$	4.6	4.9	9.8	5.1	5.0	9.5	5.3	6.3	8.6
		$\beta'$	4.6	4.8	6.2	4.3	4.5	.0	3.5	4.0	.0
	1	$\alpha'$	4.3	5.2	6.5	4.0	5.3	6.0	4.1	5.6	5.5
		$\beta'$	4.8	4.8	6.0	4.4	4.7	.0	4.1	4.0	.0
		$\alpha'$	4.8	5.7	9.8	5.0	5.4	9.3	5.3	6.5	9.3
		$\beta'$	4.6	4.5	5.9	4.6	4.5	.0	3.7	4.0	.0
1/3	1/4	$\alpha'$	.0	1.0	.1	.0	.8	.2	.1	1.1	.1
		$\beta'$	1.2	17.2	1.9	1.0	17.1	.0	1.1	16.3	.0
	1	$\alpha'$	4.4	5.1	6.0	3.8	5.0	5.4	3.1	5.6	4.3
		$\beta'$	4.6	14.5	6.1	4.0	14.4	.0	3.4	13.1	.0
		$\alpha'$	38.0	21.3	59.8	36.1	23.7	55.9	34.9	26.6	52.3
		$\beta'$	6.1	10.2	5.0	5.4	9.8	.1	4.9	9.0	.0

Note. Number of repetitions per parameter combination:  $k = 10,000$ .  $N_1/N_2 =$  ratio of sample sizes in Group 1 and 2;  $\sigma_1/\sigma_2 =$  ratio of standard deviations in Population 1 and 2;  $\delta =$  true and expected effect size (Cohen's *d* in population); SPRT = sequential probability ratio test (Hajnal's *t* test) assuming  $d = \delta$  and  $\alpha = \beta = .05$ ; GS = group sequential design with four tests, assuming  $d = \delta$  and  $\alpha = \beta = .05$ ; SBF = sequential Bayes factor design with threshold 10, assuming  $r = \sqrt{2}/2, 1, \sqrt{2}$  when  $\delta = .2, .5, \text{ and } .8$ , respectively.

This document is copyrighted by the American Psychological Association or one of its allied publishers. This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Table 4  
*Percentage of Type 2 Errors and Expected Sample Size of Hajnal's t Test, Group Sequential Test, and Sequential Bayes Factors for Random Effects*

Effect size	SPRT		GS		SBF	
	$\beta'$	ASN	$\beta'$	ASN	$\beta'$	ASN
$\delta \sim \mathcal{N}(0.2, 1)$	8.5	278	8.5	518	2.4	1042
$\delta \sim \mathcal{N}(0.5, 1)$	18.6	88	18.6	112	5.1	370
$\delta \sim \mathcal{N}(0.5, 1)$	23.6	46	23.5	52	7.0	198

Note. Number of repetitions per parameter combination:  $k = 10,000$ . SPRT = sequential probability ratio test (Hajnal's  $t$  test) assuming  $d = \mu_\delta$  and  $\alpha = \beta = .05$ ; GS = group sequential design with four tests, assuming  $d = \mu_\delta$  and  $\alpha = \beta = .05$ ; SBF = sequential Bayes factor design with threshold 10, assuming  $r = \mu_\delta$ ;  $\beta'$  = empirical error rates; ASN = average sample number ( $n_1 + n_2$ ).

assumption. However, as this assumption (fixed effect) is violated, the resulting error rates are seriously inflated. Hence, if the true effect size is in fact random rather than fixed or at least as large as expected, neither Hajnal's  $t$  test nor the GS can control the probability of a decision error.

The SBFs, on the other hand, do not put all prior weight on a single effect size but a range of effect sizes. Moreover, the prior expectation is reasonably close to the true situation in this simulation. Thus, error rates for this particular SBF are lower than for Hajnal's  $t$  test or the GS. Although a direct comparison is difficult as the designs also differ substantially in ASN, the simulation demonstrates the advantages of a diffuse prior: If the true effect is random, a diffuse prior will in general be more robust than a point prior, particularly if the scale parameter is chosen so that the prior's expected median effect size matches the true median effect size.

**Truncation Before a Decision**

**Settings.** Lastly, we address the consequences of possible misuse. One issue that might be particularly critical in practical applications of sequential tests is the risk of ending up with extremely large sample sizes. Obviously, this is not a concern in the GS design, in which an upper-bound sample size  $N_{max}$  is defined a priori. In Hajnal's  $t$  test and the SBF design, on the other hand, the final sample size is unknown. Therefore, if the sequential test has not reached a threshold at a certain point, researchers might choose to truncate it.

From a Bayesian point of view, this is not an issue. The Bayes factor is a continuous measure of evidence and its interpretation is unaffected by the stopping rule (Rouder, 2014). Hence, if the SBF procedure is terminated before reaching an a priori defined threshold, the Bayes factor at this point can still be interpreted (Schönbrodt et al., 2017). In principle, this is, of course, also possible in Hajnal's  $t$  test, as it is based on a likelihood ratio. However, this is not an option if the goal is to make decisions with a priori controlled error probabilities. Intuitively, it might seem like a reasonable strategy to truncate the sequential test when the sample size for a classical Neyman-Pearson test with corresponding error probabilities is reached, and simply switch to the fixed-sample procedure at this point. However, as only those samples will be analyzed for which the sequential procedure has not come to a

decision yet, the sampling distribution of the test statistic at this point is likely to be distorted. Any statistical inference based on it will thus be biased.

Therefore, we investigated the impact of this kind of misuse on the long-run properties of Hajnal's  $t$  test. We replicated the first simulation and truncated the process whenever the sample size reached that of a corresponding Neyman-Pearson test ( $N_{NP}$ ). A final decision was then made based on the classical  $t$  test.

**Results.** The error rates and ASNs of Hajnal's  $t$  test for a truncated sampling plan are displayed in Table 5. Additionally, it displays the proportion of replications that did not accept a hypothesis before reaching  $N_{NP}$ . Hajnal's  $t$  test consistently terminates with a sample size smaller than  $N_{NP}$  in about 90% of the cases. Hence, the risk of ending up with a larger sample is small. Nevertheless, if sampling is terminated in these cases and a decision is made based on the classical  $t$  test, the error rates are no longer fully controlled. In all cases, the nominal rate is exceeded by up to two percentage points. At the same time, the reduction in ASN compared with the open procedure is only slight. To summarize, the truncation strategy is invalid and increases the error rates beyond their nominal levels. If  $N_{NP}$  is used as the point of truncation, this increase is not dramatic but it is clearly visible and must not be ignored.

**Empirical Example**

In this section, we illustrate Hajnal's  $t$  test by applying it to a real data set. Following Schönbrodt et al. (2017), we chose open data from a replication of the retrospective gambler's fallacy (RGF) in the Many Labs Replication Project (Klein et al., 2014; <https://osf.io/ydpbf/>). The RGF, initially reported by Oppenheimer and Monin (2009), refers to people's false belief that seemingly rare outcomes are more likely to stem from a larger number of trials than seemingly common outcomes. In the experiment, participants are asked to imagine walking into a casino and observing a man rolling a die three times in a row. In the experimental condition, all dice show 6s, whereas in the control condition, two of the dice come up 6s and the third die comes up 3. Based on this scenario, participants are asked to indicate how many times they think the die had been rolled before they walked into the casino. In line with the theory, participants in the experimental group typi-

Table 5  
*Percentage of Type 1 and 2 Decision Errors and Expected Sample Size of Hajnal's t Test When Truncated at  $N_{NP}$*

$\delta$	$d$	$\alpha'/\beta'$	ASN	$N_{NP}$	% NP
.0	.2	6.9	800	1,302	9.1
	.5	6.7	134	210	10.6
	.8	6.0	54	84	10.4
.2	.2	6.5	652	1,302	8.7
.5	.5	6.4	112	210	9.9
.8	.8	5.7	48	84	10.5

Note. Number of repetitions per parameter combination:  $k = 10,000$ . Nominal error rates:  $\alpha = \beta = .05$ .  $\delta$  = true effect size (Cohen's  $d$  in population);  $d$  = expected effect size;  $\alpha'/\beta'$  = empirical error rates; ASN = average sample number ( $n_1 + n_2$ );  $N_{NP}$  = total sample size required by a Neyman-Pearson  $t$  test assuming  $\delta = d$  and  $\alpha = \beta = .05$ ; % NP = proportion of truncations at  $N_{NP}$ .

cally indicate a larger number of rolls than in the control condition. In the original study, Oppenheimer and Monin reported an effect size of Cohen's  $d = 0.69$ , 95% CI [0.16, 1.21]. In the replication study, the effect was reproduced with a total sample of  $N = 5,942$  participants, Cohen's  $d = 0.63$ , 95% CI [0.57, 0.68].

Following the *safeguard power analysis* procedure proposed by Perugini, Gallucci, and Costantini (2014), a replication of the RGF should not be based on the original effect-size estimate. Rather, one should assume, for example, the lower limit of the 80% CI of the original effect-size estimate, that is,  $d_s = 0.34$ . Thus, a replication based on a standard two-sided Neyman-Pearson  $t$  test with  $\alpha = .05$  and a power of  $1 - \beta = .95$  would require a total sample of 452 participants (Faul, Erdfelder, Buchner, & Lang, 2009). We applied Hajnal's  $t$  test with the same specifications to the data, that is,  $d = 0.34$  and  $\alpha = \beta = .05$ .

The outcome and efficiency of a sequential test depends on the sequence of observations analyzed. To avoid the impression of choosing a particular sequence, we applied the test to the data in the sequence in which they are listed in the data set. This resembles the actual application of a sequential  $t$  test, as data should be analyzed in the exact sequence in which they are sampled. Figure 3 depicts the development of the log-likelihood ratio of Hajnal's  $t$  test across the sampling process. Starting at  $N = 3$ , the test stops sampling at a total sample size of  $N = 87$  with  $LR_{87} = 19.84$ . This ratio indicates that the data are about 20 times more likely under  $\mathcal{H}_1$  than under  $\mathcal{H}_0$ , which exceeds the boundary value  $A = (1 - \beta)/\alpha = 19$ . Thus, we accept the alternative hypothesis: Participants in the RGF group indicated longer sequences ( $M = 3.55$ ,  $SD = 2.93$ ) than participants in the control group ( $M = 2.06$ ,  $SD = .98$ ), Cohen's  $d = 0.69$ , 95% CI [0.26, 1.12].<sup>6</sup> Compared with the sample size required by the standard Neyman-Pearson  $t$  test, Hajnal's  $t$  test tested the same hypothesis with the same error probabilities about 80% more efficiently.

## Discussion

Hypothesis testing is an integral part of science (Morey, Rouder, Verhagen, & Wagenmakers, 2014). It does not necessarily take the form of a dichotomous decision in favor of one of two specified hypotheses. In a Bayesian framework, for example, researchers may aim at an assessment of posterior probabilities rather than a discrete decision. Some authors even call for a shift away from hypothesis testing to inference based on estimation (Cumming, 2014; Halsey, Curran-Everett, Vowler, & Drummond, 2015; Tryon, 2016). Nevertheless, scientific discovery requires a principled, critical evaluation of whether or not a theory's predictions hold (Morey et al., 2014; Popper, 1968). For many scientists, this is represented by a binary decision to either accept or reject a hypothesis derived from the theory. As long as this decision is accompanied by an estimate of the strength of the effect, it does not conflict with the overarching aim of science to generate cumulative knowledge.

When conceiving statistical inference as decision making under uncertainty, error probabilities in statistical decisions must not be ignored, irrespective of the statistical framework used for making inferences (Lakens, 2016). Hence, employing test procedures and stopping rules that allow for error probability control is pivotal for the scientific endeavor. However, when applying statistical tests researchers also face practical constraints such as limited re-

sources. This has led to a widespread neglect of statistical power and invited a number of questionable practices, which played their part in the development of the reproducibility crisis in psychology (Bakker, van Dijk, & Wicherts, 2012; Simmons et al., 2011). Thus, in order to improve current statistical practice, sensible and efficient alternatives are needed, for example, sequential methods. Although sequential hypothesis tests have been proposed to the field of psychology in the past, their application is still surprisingly scarce in experimental research (Botella et al., 2006; Lakens, 2014; Lang, 2017).

Herein, we promote the use of the SPRT for testing precise hypotheses about mean differences between two independent groups with high efficiency and reliable control of error probabilities. The SPRT is not new. In fact, the general theory and its extensions as well as the mathematical simplifications this article builds upon have been developed more than half a century ago (Wetherill, 1975). This notwithstanding, we see three important practical and theoretical contributions of our work to psychological science.

First, in light of the ongoing reproducibility crisis, we want to introduce the SPRT to psychologists as a statistically sound and efficient alternative to the currently dominating procedure. The field is more than ever aware of the value and the need for sufficiently powered replications (Bakker et al., 2012; Lakens, 2014). Sequential methods control the probabilities of statistical decision errors while allowing for early decisions whenever the test statistic exceeds one of the boundary values, thus making optimal use of available resources. We have demonstrated the excellent properties of the SPRT for the typical two-sample  $t$  test scenario and how it is easily implemented in standard statistical software. Additionally, we created a simple, user-friendly R script to facilitate the application of the sequential tests promoted herein. Thus, the SPRT is an easy-to-apply procedure that benefits both the individual researcher and the entire field of psychology by increasing efficiency and reliably controlling error probabilities.

Second, we extended the comparison of SBFs and the GS design by Schönbrodt et al. (2017) and included the SPRT. We showed that the SPRT is more efficient not only than the GS but also than SBFs for a correctly specified hypothesis. However, it is not our intention to take a stance in the somewhat ideological quarrel between different schools of statistical inference. We merely point out the SPRT as an alternative to SBFs that (a) is more efficient when the alternative hypothesis corresponds to a point hypothesis, and (b) allows for explicit control of error probabilities. If the psychological hypothesis of interest is in fact best represented by a prior distribution rather than a point mass, we endorse the use of a correspondingly specified likelihood ratio such as the Bayes factor implanted in SBFs. In the same vein, if the research goal is to quantify evidence and assess posterior probabilities, SBFs (or generally, the Bayes factor) are the way to go. However, the standard SBF design does not allow for explicit control of error probabilities, which is a notable limitation. If error probability control is essential, the SPRT might constitute a better alternative.

<sup>6</sup> Note that this estimate of Cohen's  $d$  as well as the CI are based on the assumption of a fixed sample size and, thus, might be biased toward an overestimation of the true effect size. See the Discussion section for details.

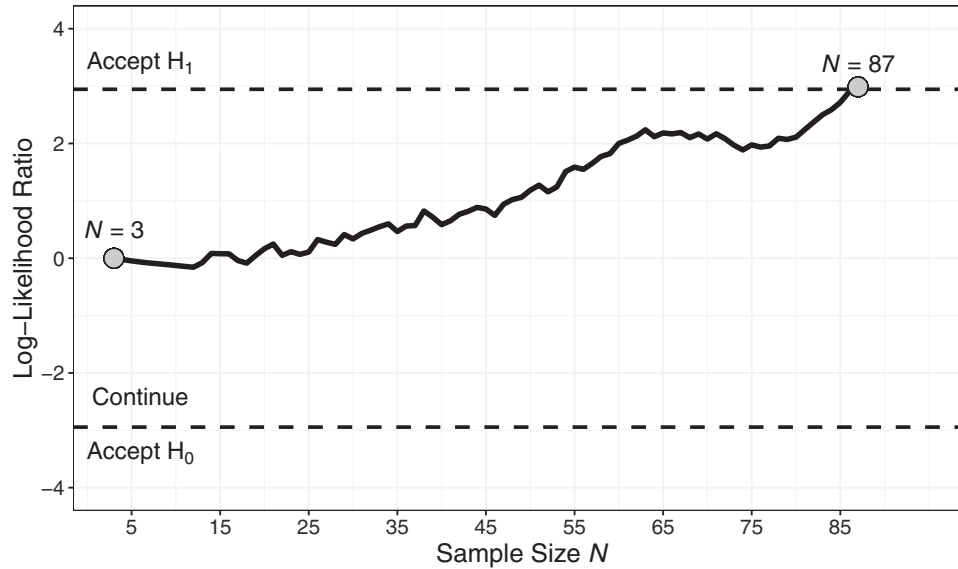


Figure 3. Development of the log-likelihood ratio for Hajnal's  $t$  test on the replication data of the retrospective gambler's fallacy (Klein et al., 2014). The test terminates sampling after  $N = 87$  observations with a decision in favor of  $\mathcal{H}_1$ . The upper and lower dashed lines represent the decision boundaries  $\ln(A)$  and  $\ln(B)$ , respectively.

Third, whereas extensive work has been done on elaborating the properties of the SPRT for simple hypotheses (Matthes, 1963; Sobel & Wald, 1949; Wald, 1947; Wald & Wolfowitz, 1948), little is known about its performance when adapted to the case of complex composite hypotheses (Cox, 1952; Köllerström & Wetherill, 1979; Wetherill, 1975). Introducing the SPRT for the two-sided two-sample  $t$  test, Hajnal (1961) stated that “there is no known method of computing the average number of observations needed for sequential tests of composite hypotheses” (p. 72). Thus, to our knowledge, our simulations constitute the first study to demonstrate the properties of Hajnal's  $t$  test for such a wide range of population scenarios and without relying on mathematical approximations to the likelihood ratio. Moreover, we examined its robustness against a number of violations of its basic assumptions and compared this with the robustness of SBFs and the GS design. To summarize our results, in a balanced design and when the effect size is not grossly misspecified, Hajnal's  $t$  test is highly efficient and quite robust even under conditions of non-normality or heteroscedasticity.

### Limitations

There are some possible limitations of our work that apply to sequential procedures in general, whereas others are specific for the test we promote in this article. First, some critics might object that the SPRT requires a precise specification of both the null and the alternative hypothesis (i.e., a precise prediction of the effect size). Ideally, this prediction follows from an underlying theory; however, it is frequently argued that researchers do not have realistic effect-size assumptions (Gelman & Carlin, 2014; Perugini et al., 2014). If there is no information in the literature such that an effect-size estimate could be based on a review or meta-analysis, this may indeed seem like a severe drawback. However, it is important to keep in mind that the effect-size assumption under  $\mathcal{H}_1$

is not necessarily an attempt to guess the true effect that underlies the data. Alternatively, it can be seen as specification of an effect that the researcher “deem[s] worthy of detecting” (Schulz & Grimes, 2005, p. 1350). Thus, the need to specify a precise hypothesis should not be considered detrimental. After all, a hypothesis test is, by definition, the test of a prediction—why would we demand it to work without specifying one? NHST is an inglorious example of the critical consequences of employing a test without specifying a precise alternative to the null hypothesis.

It is true, however, that the SPRT will be less efficient or may lead to wrong decisions more often when the effect size is grossly under- or overspecified, respectively. At the cost of efficiency, the SBF design is more robust against such misspecifications to a certain extent. However, this does by no means imply that one need not define a sensible statistical hypothesis: If the prior allocates undue mass to effect sizes that differ substantially from the true effect, the resulting test procedure will also perform poorly in terms of asymptotic error rates and efficiency. In sum, sensible hypothesis tests require reasonable and precise statistical hypotheses; the more precise a hypothesis, the more critical and efficient is its test (Stefan et al., 2019).

Second, the SPRT is an open procedure, that is, it requires sampling until a decision is made. It cannot be ruled out a priori that the data do not yield strong evidence in favor of any hypothesis such that the test goes on for thousands of observations. However, our results indicate that the risk of extremely large sample sizes in Hajnal's  $t$  test is small, although such events are possible in principle. Obviously, this is a potential risk in any open sequential design, SPRT and SBFs alike. Next to the GS design, there have been suggestions in the literature to modify sequential procedures such that they definitely terminate at or before a certain sample size  $N_{max}$  (Armitage, 1957; T. W. Anderson, 1960). However, as the comparison with the GS demonstrated, these restricted

tests are not optimal, that is, they either are less powerful or come with higher ASNs than open sequential designs (Wetherill, 1975).

Because the SPRT is based on a likelihood ratio, like SBFs, it is possible to define an  $N_{max}$  at which sampling terminates even if no boundary is reached. One could then report the likelihood ratio at  $N_{max}$ . However, such a procedure cannot be used for dichotomous decisions with controlled error probabilities, because error probabilities would be larger to an unknown degree than those of the open sequential test. Specifically, the smaller  $N_{max}$  at the point of termination, the higher the extent to which the error probabilities of the truncated test exceed those of the open test (Wetherill, 1975). In the same vein, we demonstrated with our simulations that it would be ill-advised to administer a standard fixed-sample test after a sequential test failed to find a decision within the sample size defined by an a priori power analysis. Hence, it is important to either continue until a boundary is reached or terminate without a definite decision and report the observed likelihood ratio only.

So far, our discussion focused only on the properties of sequential designs as efficient and accurate procedures to decide between two statistical hypotheses. As elucidated at the outset of the discussion, deciding in favor of a hypothesis is not the only means of statistical inference. It merely represents the process of accepting the data as corroboration or refutation of a prediction of interest. However, the scope of information in the data goes beyond this binary decision and should be conveyed in the form of effect-size estimates. Herein, we did not explicitly address the issue of effect-size estimation following the sequential procedure because it is not unique to the SPRT and has been addressed before (e.g., Emerson & Fleming, 1990; Fan, DeMets, & Lan, 2004; Goodman, 2007; Mueller, Montori, Bassler, Koenig, & Guyatt, 2007; Schönbrodt & Wagenmakers, 2018; Schönbrodt et al., 2017; Stallard, Todd, & Whitehead, 2008; Whitehead, 1986; Zhang et al., 2012).

The difficulty of estimation following a sequential test resulting in acceptance of  $\mathcal{H}_1$  arises from the fact that the evidence in the data, which is reflected in the effect-size estimate, determines the sample size. Strong evidence for  $\mathcal{H}_1$  will result in early stopping, whereas weaker evidence will lead to larger samples. Hence, the sampling distribution of effect-size estimates will be distorted considerably, with small samples systematically overestimating and large samples systematically underestimating the true effect of interest (Whitehead, 1986; Zhang et al., 2012).

However, a closer look reveals that this apparent drawback is not as serious as it may seem (Goodman, 2007; Schönbrodt & Wagenmakers, 2018): The overestimation of effect sizes by only considering early terminations at  $\mathcal{H}_1$  is comparable with the overestimation of effect sizes caused by publication bias (see Ulrich, Miller, & Erdfelder, 2018). That is, it is based on a loss of information rather than the sequential nature of the test procedure itself. When aggregating across early and late terminations, the bias—although it remains—is reduced and might be considered negligible (Schönbrodt & Wagenmakers, 2018). Moreover, the SPRT should be less prone to publication bias than NHST because it allows for acceptance of both hypotheses. Hence, meta-analytical effect-size estimates taking into account sample sizes and estimates from both early and late terminations in favor of  $\mathcal{H}_1$  or  $\mathcal{H}_0$  will basically be unbiased (see Schönbrodt et al., 2017).

## Practical Recommendations

The SPRT we promote in this article can easily be set up with any statistical software in which the probability density functions of  $t$  or  $F$  are provided or can be implemented. A workable, user-friendly R script to perform Hajnal's  $t$  test can be downloaded from the Open Science Framework (<https://osf.io/4zub2/>). Herein, we explicitly addressed the case of testing two-sided hypotheses for two independent groups. The script additionally can be used to perform a sequential  $t$  test for one-sided hypotheses as well as hypotheses about a single or two dependent groups. Note, however, that the expected sample sizes observed in our simulations apply only to the two-sided two-sample scenario (Hajnal's  $t$  test). Smaller ASNs can be expected for one-sided hypotheses and dependent observations.

As noted earlier, there are different ways to specify a sensible alternative hypothesis. Ideally, one has a precise prediction implied by a psychological theory. However, we acknowledge that this is not always the case. If an effect-size assumption is based on previous estimates in the literature, it makes sense to take the uncertainty of these estimates into account and assume a lower-bound effect size to ensure a sufficiently powered test (Perugini et al., 2014). Similarly, in total absence of any information or precise prediction, one should specify a minimum relevant effect  $d_{min}$  to obtain a power of at least  $1 - \beta$  for the SPRT to detect an effect  $\delta \geq d_{min}$ . Note, however, that a conservative effect-size assumption will result in a less efficient test.

To make sure that error rates as specified by  $\alpha$  and  $\beta$  are not exceeded, the data need to be analyzed in the sequence in which they have been sampled. This must be continued until the inequality  $B < LR < A$  is violated, resulting in a decision for one of the two hypotheses of interest. Hajnal's  $t$  test does not require pairwise sampling in general (H. Lee & Fung, 1980). Participants can be randomly allocated to a group and the data can be analyzed after each additional observation irrespective of the relative group sizes, as long as there are at least three observations in total and at least one in each group. However, we strongly recommend a balanced design as this will increase the test's efficiency and robustness in case of heteroscedasticity (see the Robustness of SPRT, GS, and SBFs section).

In sum, when testing hypotheses with the SPRT one should adhere to the following simple steps:

1. Specify the statistical hypotheses (e.g., the to-be-detected minimal effect size  $d$ ) and the desired upper bounds to the error probabilities of the test ( $\alpha$ ,  $\beta$ ) before the sampling process. Do not alter these specifications during the sampling process in response to the data observed.
2. Analyze the data in the sequence in which they have been sampled. This sequence must not be altered to obtain a specific result (e.g., by dropping unwanted observations). Observations may be added and analyzed in groups rather than separately. However, this may result in a decrease of error probabilities and, correspondingly, efficiency.
3. Continue sampling as long as  $\beta/(1 - \alpha) < LR < (1 - \beta)/\alpha$  and terminate as soon as this inequality is violated,



resulting in a decision in favor of  $\mathcal{H}_0$  if  $LR \leq \beta/(1 - \alpha)$  or  $\mathcal{H}_1$  if  $LR \geq (1 - \beta)/\alpha$ .

### Conclusion

Sequential analyses are useful tools to conduct sufficiently powered hypothesis tests with minimal costs in terms of time and observations needed. Particularly in light of the ongoing reproducibility crisis, these are highly desirable features that could benefit both individual researchers and the entire field of psychological science (Lakens, 2014). We showed that the SPRT is not only easily applied to the common  $t$  test scenario but also more efficient than other common sequential designs. Additionally, the SPRT allows for specifying reliable upper bounds to decision error probabilities.

We do not promote the SPRT as the single optimal inference procedure for all situations. After all, statistics is not a single tool that fits all problems but a toolbox that contains several procedures suited for different situations. Depending on the aim of the researcher and the problem at hand, some research questions may better be approached using a fixed-sample design, and others by a different sequential design such as SBFs, GS, or adaptive designs (Lakens & Evers, 2014). With this article, we hope to expand the scope of psychologists' statistical toolboxes by proposing the SPRT as an efficient alternative to conventional methods of controlling statistical decision errors.

### References

- Anderson, K. (2014). *gsDesign: Group sequential design*. Retrieved from <http://CRAN.R-project.org/package=gsDesign>
- Anderson, T. W. (1960). A modification of the sequential probability ratio test to reduce the sample size. *The Annals of Mathematical Statistics*, 31, 165–197. <http://dx.doi.org/10.1214/aoms/1177705996>
- Anscombe, F. J. (1954). Fixed-sample-size analysis of sequential observations. *Biometrics*, 10, 89–100. <http://dx.doi.org/10.2307/3001665>
- Armitage, P. (1947). Some sequential tests of student's hypothesis. *Supplement to the Journal of the Royal Statistical Society*, 9, 250–263. <http://dx.doi.org/10.2307/2984117>
- Armitage, P. (1957). Restricted sequential procedures. *Biometrika*, 44, 9–26. <http://dx.doi.org/10.2307/2333237>
- Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., . . . Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, 27, 108–119. <http://dx.doi.org/10.1002/per.1919>
- Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, 66, 423–437. <http://dx.doi.org/10.1037/h0020412>
- Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, 7, 543–554. <http://dx.doi.org/10.1177/1745691612459060>
- Barnard, G. A. (1949). Statistical inference. *Journal of the Royal Statistical Society Series B (Methodological)*, 11, 115–149. <http://dx.doi.org/10.1007/978-1-4613-8505-9>
- Berger, J. O. (2003). Could Fisher, Jeffreys, and Neyman have agreed on testing? *Statistical Science*, 18, 1–32. <http://dx.doi.org/10.1214/ss/1056397485>
- Botella, J., Ximénez, C., Revuelta, J., & Suero, M. (2006). Optimization of sample size in controlled experiments: The CLAST rule. *Behavior Research Methods*, 38, 65–76. <http://dx.doi.org/10.3758/BF03192751>
- Bredenkamp, J. (1972). *Der Signifikanztest in Der Psychologischen Forschung* [The test of significance in psychological research]. Darmstadt, Germany: Akademische Verlagsgesellschaft.
- Clopper, C. J., & Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26, 404–413. <http://dx.doi.org/10.2307/2331986>
- Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology*, 65, 145–153. <http://dx.doi.org/10.1037/h0045186>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Cox, D. R. (1952). Sequential tests for composite hypotheses. *Mathematical Proceedings of the Cambridge Philosophical Society*, 48, 290–299. <http://dx.doi.org/10.1017/S030500410002764X>
- Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, 25, 7–29. <http://dx.doi.org/10.1177/0956797613504966>
- Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, 6, 274–290. <http://dx.doi.org/10.1177/1745691611406920>
- Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, 6, 621. <http://dx.doi.org/10.3389/fpsyg.2015.00621>
- Emerson, S. S., & Fleming, T. R. (1990). Parameter estimation following group sequential hypothesis testing. *Biometrika*, 77, 875–892. <http://dx.doi.org/10.2307/2337110>
- Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers*, 28, 1–11. <http://dx.doi.org/10.3758/BF03203630>
- Etz, A. (2018). Introduction to the concept of likelihood and its applications. *Advances in Methods and Practices in Psychological Science*, 1, 60–69. <http://dx.doi.org/10.1177/2515245917744314>
- 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, 505–530. <http://dx.doi.org/10.1081/BIP-120037195>
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G\*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41, 1149–1160. <http://dx.doi.org/10.3758/BRM.41.4.1149>
- Gelman, A. (2016). Commentary on “Crisis in science? Or crisis in statistics! Mixed messages in statistics with impact on science.” *Journal of Statistical Research*, 48–50, 11–12.
- Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, 9, 641–651. <http://dx.doi.org/10.1177/1745691614551642>
- Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences* (pp. 311–339). Hillsdale, NJ: Erlbaum.
- Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, 33, 587–606. <http://dx.doi.org/10.1016/j.socec.2004.09.033>
- Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on “Estimating the reproducibility of psychological science.” *Science*, 351, 1037. <http://dx.doi.org/10.1126/science.aad7243>
- Goodman, S. N. (1993). P values, hypothesis tests, and likelihood: Implications for epidemiology of a neglected historical debate. *American Journal of Epidemiology*, 137, 485–496. <http://dx.doi.org/10.1093/oxfordjournals.aje.a116700>
- Goodman, S. N. (2007). Stopping at nothing? Some dilemmas of data monitoring in clinical trials. *Annals of Internal Medicine*, 146, 882–887. <http://dx.doi.org/10.7326/0003-4819-146-12-200706190-00010>
- Hajnal, J. (1961). A two-sample sequential t-test. *Biometrika*, 48, 65–75. <http://dx.doi.org/10.2307/2333131>
- Halsey, L. G., Curran-Everett, D., Vowler, S. L., & Drummond, G. B. (2015). The fickle p value generates irreproducible results. *Nature Methods*, 12, 179–185. <http://dx.doi.org/10.1038/nmeth.3288>

- Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. *Mathematical Proceedings of the Cambridge Philosophical Society*, 31, 203–222. <http://dx.doi.org/10.1017/S030500410001330X>
- Jeffreys, H. (1961). *Theory of probability*. New York, MU: Oxford University Press.
- Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90, 773–795. <http://dx.doi.org/10.1080/01621459.1995.10476572>
- Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B. J., Bahník, Š., Bernstein, M. J., . . . Nosek, B. A. (2014). Investigating variation in replicability: A “many labs” replication project. *Social Psychology*, 45, 142–152. <http://dx.doi.org/10.1027/1864-9335/a000178>
- Köllerström, J., & Wetherill, G. B. (1979). SPRT's for the normal correlation coefficient. *Journal of the American Statistical Association*, 74, 815–821. <http://dx.doi.org/10.2307/2286405>
- Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, 44, 701–710. <http://dx.doi.org/10.1002/ejsp.2023>
- Lakens, D. (2016). Dance of the Bayes factors [Blog post]. Retrieved from <http://daniellakens.blogspot.de/2016/07/dance-of-bayes-factors.html>
- 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, 278–292. <http://dx.doi.org/10.1177/1745691614528520>
- Lang, A.-G. (2017). Is intermediately inspecting statistical data necessarily a bad research practice? *The Quantitative Methods for Psychology*, 13, 127–140. <http://dx.doi.org/10.20982/tqmp.13.2.p127>
- Lee, H., & Fung, K. Y. (1980). A Monte Carlo study on the robustness of the two-sample sequential  $t$ -test. *Journal of Statistical Computation and Simulation*, 10, 297–307. <http://dx.doi.org/10.1080/00949658.008810377>
- Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. New York, NY: Cambridge University Press.
- Matthes, T. K. (1963). On the optimality of sequential probability ratio tests. *The Annals of Mathematical Statistics*, 34, 18–21. <http://dx.doi.org/10.1214/aoms/1177704239>
- Matzke, D., Nieuwenhuis, S., van Rijn, H., Slagter, H. A., van der Molen, M. W., & Wagenmakers, E.-J. (2015). The effect of horizontal eye movements on free recall: A preregistered adversarial collaboration. *Journal of Experimental Psychology: General*, 144, e1–e15. <http://dx.doi.org/10.1037/xge0000038>
- Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does “failure to replicate” really mean? *American Psychologist*, 70, 487–498. <http://dx.doi.org/10.1037/a0039400>
- Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods*, 16, 406–419. <http://dx.doi.org/10.1037/a0024377>
- Morey, R. D., & Rouder, J. N. (2015). *BayesFactor: Computation of Bayes factors for common designs*. Retrieved from <https://CRAN.R-project.org/package=BayesFactor>
- Morey, R. D., Rouder, J. N., Verhagen, J., & Wagenmakers, E.-J. (2014). Why hypothesis tests are essential for psychological science: A comment on Cumming (2014). *Psychological Science*, 25, 1289–1290. <http://dx.doi.org/10.1177/0956797614525969>
- Mueller, P. S., Montori, V. M., Bassler, D., Koenig, B. A., & Guyatt, G. H. (2007). Ethical issues in stopping randomized trials early because of apparent benefit. *Annals of Internal Medicine*, 146, 878–882. <http://dx.doi.org/10.7326/0003-4819-146-12-200706190-00009>
- 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. <http://dx.doi.org/10.1098/rsta.1933.0009>
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349, aac4716. <http://dx.doi.org/10.1126/science.aac4716>
- Oppenheimer, D. M., & Monin, B. (2009). The retrospective gambler's fallacy: Unlikely events, constructing the past, and multiple universes. *Judgment and Decision Making*, 4, 326–334.
- Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, 7, 528–530. <http://dx.doi.org/10.1177/1745691612465253>
- Perugini, M., Gallucci, M., & Costantini, G. (2014). Safeguard power as a protection against imprecise power estimates. *Perspectives on Psychological Science*, 9, 319–332. <http://dx.doi.org/10.1177/1745691614528519>
- Popper, K. R. (1968). *The logic of scientific discovery* (3rd ed.). London, UK: Hutchinson.
- Proschan, M. A., Lan, K. K. G., & Wittes, J. T. (2006). *Statistical monitoring of clinical trials: A unified approach*. New York, NY: Springer.
- Psychonomic Society. (2012). *Psychonomic Society statistical guidelines*. Retrieved from <http://www.psychonomic.org/page/statisticalguideline>
- R Core Team. (2017). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <http://www.R-project.org/>
- Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, 21, 301–308. <http://dx.doi.org/10.3758/s13423-014-0595-4>
- 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, 225–237. <http://dx.doi.org/10.3758/PBR.16.2.225>
- Rushton, S. (1950). On a sequential  $t$ -test. *Biometrika*, 37, 326–333. <http://dx.doi.org/10.2307/2332385>
- Rushton, S. (1952). On a two-sided sequential  $t$ -test. *Biometrika*, 39, 302. <http://dx.doi.org/10.2307/2334026>
- Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, 25, 128–142. <http://dx.doi.org/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, 322–339. <http://dx.doi.org/10.1037/met0000061>
- Schulz, K. F., & Grimes, D. A. (2005). Sample size calculations in randomised trials: Mandatory and mystical. *The Lancet*, 365, 1348–1353. [http://dx.doi.org/10.1016/S0140-6736\(05\)61034-3](http://dx.doi.org/10.1016/S0140-6736(05)61034-3)
- Sedlmeier, P. (1996). Jenseits des Signifikanztest-Rituals: Ergänzungen und Alternativen [Beyond the ritual of significance testing: Alternative and supplementary methods]. *Methods of Psychological Research Online*, 1, 41–63. Retrieved from <https://www.dgps.de/fachgruppen/methoden/mpr-online/>
- Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, 105, 309–316. <http://dx.doi.org/10.1037/0033-2909.105.2.309>
- 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. <http://dx.doi.org/10.1177/0956797611417632>
- Sobel, M., & Wald, A. (1949). A sequential decision procedure for choosing one of three hypotheses concerning the unknown mean of a normal distribution. *The Annals of Mathematical Statistics*, 20, 502–522. <http://dx.doi.org/10.1214/aoms/1177729944>
- Stallard, N., Todd, S., & Whitehead, J. (2008). Estimation following selection of the largest of two normal means. *Journal of Statistical*

- Planning and Inference*, 138, 1629–1638. <http://dx.doi.org/10.1016/j.jspi.2007.05.045>
- Stefan, A. M., Gronau, Q. F., Schönbrodt, F. D., & Wagenmakers, E.-J. (2019). A tutorial on Bayes factor design analysis using an informed prior. *Behavior Research Methods*, 51, 1042–1058. <http://dx.doi.org/10.3758/s13428-018-01189-8>
- Tryon, W. W. (2016). Replication is about effect size: Comment on Maxwell, Lau, and Howard (2015). *American Psychologist*, 71, 236–237. <http://dx.doi.org/10.1037/a0040191>
- Ulrich, R., Miller, J., & Erdfelder, E. (2018). Effect size estimation from t-statistics in the presence of publication bias: A brief review of existing approaches with some extensions. *Zeitschrift für Psychologie*, 226, 56–80. <http://dx.doi.org/10.1027/2151-2604/a000319>
- Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, 14, 779–804. <http://dx.doi.org/10.3758/BF03194105>
- Wagenmakers, E.-J., Beek, T. F., Rotteveel, M., Gierholz, A., Matzke, D., Steingroever, H., . . . Pinto, Y. (2015). Turning the hands of time again: A purely confirmatory replication study and a Bayesian analysis. *Frontiers in Psychology*, 6, 494. <http://dx.doi.org/10.3389/fpsyg.2015.00494>
- Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16, 117–186.
- Wald, A. (1947). *Sequential analysis*. New York, NY: Wiley.
- Wald, A., & Wolfowitz, J. (1948). Optimum character of the sequential probability ratio test. *The Annals of Mathematical Statistics*, 19, 326–339. <http://dx.doi.org/10.1214/aoms/1177730197>
- Wetherill, G. B. (1975). *Sequential methods in statistics* (2nd ed.). London, UK: Chapman and Hall.
- Whitehead, J. (1986). On the bias of maximum likelihood estimation following a sequential test. *Biometrika*, 73, 573–581. <http://dx.doi.org/10.1093/biomet/73.3.573>
- Wrinch, D., & Jeffreys, H. (1921). On certain fundamental principles of scientific inquiry. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 42, 369–390. <http://dx.doi.org/10.1080/14786442108633773>
- Zellner, A., & Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. *Trabajos de Estadística Y de Investigación Operativa*, 31, 585–603. <http://dx.doi.org/10.1007/BF02888369>
- Zhang, J. J., Blumenthal, G. M., He, K., Tang, S., Cortazar, P., & Sridhara, R. (2012). Overestimation of the effect size in group sequential trials. *Clinical Cancer Research*, 18, 4872–4876. <http://dx.doi.org/10.1158/1078-0432.CCR-11-3118>

This document is copyrighted by the American Psychological Association or one of its allied publishers. This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.

Appendix  
Average Sample Numbers

Table A1  
Expected Sample Sizes (and Quantiles) of Hajnal's  $t$  Test, Group Sequential Test, and Sequential Bayes Factors

$d$	$r$	Hajnal's $t$ test						Group sequential test						Sequential Bayes factors		
		$\alpha = 1\%$		$\alpha = 5\%$		$\alpha = 1\%$		$\alpha = 5\%$		$\alpha = 1\%$		$\alpha = 5\%$		BF = 5	BF = 10	BF = 30
		$\beta = 5\%$	$\beta = 10\%$	$\beta = 5\%$	$\beta = 10\%$	$\beta = 5\%$	$\beta = 10\%$	$\beta = 5\%$	$\beta = 10\%$	$\beta = 5\%$	$\beta = 10\%$	$\beta = 5\%$	$\beta = 10\%$			
$\delta = 0$																
.2	$\sqrt{2}/2$	888	692	832	656	1,214	1,006	1,040	834	174	834	834	174	834	834	8160
		(734, 982, 1,722)	(560, 764, 1,360)	(704, 920, 1,512)	(544, 732, 1,222)	(964, 1,446, 1,928)	(806, 1,208, 1,612)	(1,062, 1,062, 1,416)	(862, 862, 1,148)	(136, 186, 396)	(594, 844, 1,980)	(5524, 7,878, 19,588)				
.5	1	146	116	138	110	198	166	170	138	92	426	4152	92	426	426	4,152
		(122, 162, 278)	(94, 128, 230)	(116, 152, 250)	(90, 122, 208)	(158, 236, 314)	(132, 196, 262)	(172, 172, 230)	(142, 142, 188)	(70, 96, 210)	(300, 418, 1,010)	(2,782, 3,962, 10,202)				
.8	$\sqrt{2}$	60	48	56	46	80	68	68	56	48	216	2,070	48	216	216	2,070
		(48, 66, 116)	(38, 54, 98)	(48, 62, 102)	(38, 50, 88)	(64, 94, 126)	(54, 80, 106)	(70, 70, 92)	(56, 56, 74)	(36, 50, 112)	(152, 214, 504)	(1,386, 1,986, 4,882)				
$\delta = .2$																
.2	$\sqrt{2}/2$	998	930	690	626	1,278	1,150	932	808	236	770	1,120	236	770	770	1,120
		(870, 1,254, 2,072)	(812, 1,164, 1,890)	(592, 880, 1,538)	(534, 788, 1,350)	(1,446, 1,446, 1,928)	(1,208, 1,612, 1,612)	(1,062, 1,062, 1,416)	(862, 1,148, 1,148)	(150, 276, 710)	(650, 1,078, 1,868)	(994, 1,546, 2,518)				
.5	1	232	178	178	138	236	194	184	148	126	608	1,184	126	608	608	1,184
		(160, 286, 592)	(118, 218, 474)	(134, 212, 412)	(102, 160, 320)	(236, 314, 314)	(196, 262, 262)	(172, 230, 230)	(142, 188, 188)	(74, 128, 390)	(450, 848, 1,604)	(1,070, 1,636, 2,624)				
.8	$\sqrt{2}$	76	60	66	52	88	72	72	58	62	384	1,236	62	384	384	1,236
		(54, 84, 182)	(42, 66, 144)	(50, 76, 144)	(38, 58, 110)	(94, 94, 126)	(80, 80, 106)	(70, 92, 92)	(56, 74, 74)	(38, 60, 178)	(204, 484, 1,254)	(1,118, 1,696, 2,688)				
$\delta = .5$																
.2	$\sqrt{2}/2$	276	272	194	190	496	436	378	328	100	132	172	100	132	132	172
		(262, 320, 428)	(260, 316, 420)	(180, 226, 320)	(178, 224, 314)	(482, 482, 482)	(404, 404, 806)	(354, 354, 708)	(288, 288, 574)	(84, 140, 250)	(112, 180, 314)	(152, 232, 374)				
.5	1	168	158	120	110	208	186	150	132	88	134	176	88	134	134	176
		(146, 210, 344)	(138, 198, 322)	(100, 150, 258)	(92, 136, 228)	(236, 236, 314)	(196, 214, 262)	(172, 172, 230)	(142, 142, 188)	(70, 120, 216)	(114, 186, 320)	(156, 240, 380)				
.8	$\sqrt{2}$	112	92	78	62	102	86	74	60	64	138	184	64	138	138	184
		(86, 142, 266)	(70, 116, 224)	(60, 96, 180)	(48, 76, 144)	(94, 126, 126)	(80, 106, 106)	(70, 92, 92)	(56, 74, 74)	(44, 82, 170)	(120, 192, 324)	(164, 248, 404)				
$\delta = .8$																
.2	$\sqrt{2}/2$	166	164	116	114	482	404	354	290	44	56	72	44	56	56	72
		(160, 184, 226)	(158, 182, 222)	(112, 130, 166)	(110, 128, 162)	(482, 482, 482)	(404, 404, 404)	(354, 354, 354)	(288, 288, 288)	(36, 60, 108)	(48, 74, 124)	(64, 94, 148)				
.5	1	86	86	62	60	124	114	94	86	44	56	72	44	56	56	72
		(80, 102, 148)	(78, 100, 144)	(54, 72, 112)	(54, 72, 110)	(158, 158, 158)	(132, 132, 196)	(116, 116, 172)	(94, 94, 142)	(36, 60, 106)	(48, 74, 126)	(64, 94, 150)				
.8	$\sqrt{2}$	72	68	50	48	82	74	60	52	42	58	74	42	58	58	74
		(62, 88, 142)	(58, 82, 130)	(42, 62, 106)	(40, 58, 96)	(94, 94, 126)	(80, 80, 106)	(70, 70, 92)	(56, 56, 74)	(34, 56, 98)	(48, 78, 130)	(66, 98, 154)				

Note. Depicted are expected total sample sizes ( $n_1 + n_2$ ) and the 50th, 75th, and 95th quantile in parentheses. Note that the  $d$  and  $r$  metrics indicate fundamentally different effect-size expectations, although they are often assigned the same verbal labels for "small" ( $d = .2, r = \sqrt{2}/2$ ), "medium" ( $d = .5, r = 1$ ), and "large" effects ( $d = .8, r = \sqrt{2}$ ). The group sequential test comprised three interim and one final test.  $d$  = expected effect size according to  $\mathcal{H}_1$  in Hajnal's  $t$  test and group sequential test (Cohen's  $d$ );  $r$  = scale parameter of Cauchy prior (= expected median absolute effect size) according to  $\mathcal{H}_1$  in sequential Bayes factors; BF = threshold Bayes factor;  $\delta$  = true population effect size.

(Appendices continue)

Table A2  
*Expected Sample Sizes of Hajnal's t Test, Group Sequential Test, and Sequential Bayes Factors Under Conditions of Non-Normality*

Distribution	$\gamma$	$s$	$k$	True state	$\delta = .2$			$\delta = .5$			$\delta = .8$		
					SPRT	GS	SBF	SPRT	GS	SBF	SPRT	GS	SBF
Normal		.0	3.0	$\mathcal{H}_0$	832	1,040	834	138	170	426	56	68	216
				$\mathcal{H}_1$	690	932	770	120	150	134	50	60	58
Mixture	.9	.8	6.0	$\mathcal{H}_0$	834	1,036	832	140	170	432	58	68	222
				$\mathcal{H}_1$	674	914	748	114	144	130	48	58	54
	.7	.9	4.4	$\mathcal{H}_0$	830	1,018	822	140	166	434	58	68	218
				$\mathcal{H}_1$	626	872	698	108	140	118	46	54	48
	.5	.7	3.4	$\mathcal{H}_0$	794	994	732	138	162	384	58	66	204
				$\mathcal{H}_1$	562	832	612	96	134	104	42	52	46
Log-normal		6.2	116.9	$\mathcal{H}_0$	860	1,040	886	146	172	454	60	70	232
				$\mathcal{H}_1$	642	892	682	94	132	96	36	48	36

Note. Depicted are expected total sample sizes ( $n_1 + n_2$ ). The first two rows display results from the first simulation for normally distributed data (see Table A1, Columns 3, 7, and 10). Number of repetitions per parameter combination:  $k = 10,000$ .  $\gamma$  = mixture probability;  $s$  = skewness;  $k$  = kurtosis;  $\delta$  = true and expected effect size (Cohen's  $d$  in population); SPRT = sequential probability ratio test (Hajnal's  $t$  test) assuming  $d = \delta$  and  $\alpha = \beta = .05$ ; GS = group sequential design with four tests, assuming  $d = \delta$  and  $\alpha = \beta = .05$ ; SBF = sequential Bayes factor design with threshold 10, assuming  $r = \sqrt{2}/2, 1, \sqrt{2}$  when  $\delta = .2, .5,$  and  $.8$ , respectively;  $\mathcal{H}_0, \mathcal{H}_1$  = true state underlying data generation.

Table A3  
*Expected Sample Sizes of Hajnal's t Test, Group Sequential Test, and Sequential Bayes Factors Under Conditions of Heteroscedasticity*

$N_1/N_2$	$\sigma_1/\sigma_2$	True state	$\delta = .2$			$\delta = .5$			$\delta = .8$		
			SPRT	GS	SBF	SPRT	GS	SBF	SPRT	GS	SBF
1	1/4	$\mathcal{H}_0$	832	1,028	797	137	166	410	56	67	204
		$\mathcal{H}_1$	683	924	738	116	148	129	49	58	54
	1	$\mathcal{H}_0$	834	1,024	820	139	166	416	57	67	212
		$\mathcal{H}_1$	688	926	770	119	149	133	50	59	57
	4	$\mathcal{H}_0$	831	1,023	784	137	167	408	56	67	210
		$\mathcal{H}_1$	687	928	744	115	149	128	49	58	53
1/3	1/4	$\mathcal{H}_0$	993	935	914	164	152	454	67	62	232
		$\mathcal{H}_1$	1,471	1,213	1,871	240	195	282	99	77	112
	1	$\mathcal{H}_0$	1,115	1,027	1,112	185	166	573	77	68	292
		$\mathcal{H}_1$	916	1,025	1,044	157	166	186	68	65	78
	4	$\mathcal{H}_0$	875	1,025	507	152	164	286	65	65	164
		$\mathcal{H}_1$	533	792	321	95	127	81	42	49	39

Note. Depicted are expected total sample sizes ( $n_1 + n_2$ ). Number of repetitions per parameter combination:  $k = 10,000$ .  $N_1/N_2$  = ratio of sample sizes in Group 1 and 2;  $\sigma_1/\sigma_2$  = ratio of standard deviations in population 1 and 2;  $\delta$  = true and expected effect size (Cohen's  $d$  in population); SPRT = sequential probability ratio test (Hajnal's  $t$  test) assuming  $d = \delta$  and  $\alpha = \beta = .05$ ; GS = group sequential design with four tests, assuming  $d = \delta$  and  $\alpha = \beta = .05$ ; SBF = sequential Bayes factor design with threshold 10, assuming  $r = \sqrt{2}/2, 1, \sqrt{2}$  when  $\delta = .2, .5,$  and  $.8$ , respectively;  $\mathcal{H}_0, \mathcal{H}_1$  = true state underlying data generation.

Received December 18, 2018

Revision received June 3, 2019

Accepted June 20, 2019 ■