Waldian t Tests: Sequential Bayesian t Tests with Controlled Error Probabilities

Martin Schnuerch[1], Daniel W. Heck[2], & Edgar Erdfelder[1]

[1] University of Mannheim
[2] University of Marburg

Abstract

Bayesian t tests have become increasingly popular alternatives to null-hypothesis significance testing (NHST) in psychological research. In contrast to NHST, they allow for the quantification of evidence in favor of the null hypothesis and for optional stopping. A major drawback of Bayesian t tests, however, is that error probabilities of statistical decisions remain uncontrolled. Previous approaches in the literature to remedy this problem require time-consuming simulations to calibrate decision thresholds. In this article, we propose a sequential probability ratio test that combines Bayesian t tests with simple decision criteria developed by Abraham Wald in 1947. We discuss this sequential procedure, which we call Waldian t test, in the context of three recently proposed specifications of Bayesian t tests. We show that Waldian t tests reliably control frequentist error probabilities, with the nominal Type I and Type II error probabilities serving as upper bounds to the actual error rates. At the same time, the key idea of Bayesian t tests (i.e, assuming a prior distribution for the effect size under the alternative hypothesis) is preserved. Thus, Waldian t tests are fully justified from both a frequentist and a Bayesian point of view. We highlight the relationship between frequentist and Bayesian error probabilities and critically discuss the implications of conventional stopping criteria for sequential Bayesian t tests. Finally, we provide a user-friendly web application that implements the proposed procedure for interested researchers.

Keywords: Bayesian t tests, Bayes factors, statistical error probabilities, sequential tests, sequential probability ratio test

A key component of empirical science is the critical evaluation of hypotheses in light of data. Whereas there is little controversy about this statement, the particular means by which hypotheses should be tested has been the subject of many heated debates. *Null-hypothesis significance testing* (NHST), the dominant procedure in psychology, has been harshly criticized for decades (e.g., Bakan, 1966; Bredenkamp, 1972; Cohen, 1994; Gelman, 2016; Gigerenzer, 1993, 2004; Rozeboom, 1960; Wagenmakers, 2007). Some of these criticisms aim at misinterpretations and misuse of p-value based procedures. In response, there have been suggestions to improve their application by stressing the importance of considering effect sizes and power analysis within the Neyman-Pearson framework (Cohen, 1962, 1994; Erdfelder et al., 1996) and of correctly interpreting p values (Lakens, 2021). Other critics, in contrast, have called into question the adequacy of p values and frequentist methods for hypothesis testing in general. Consequently, there have been several calls to abandon frequentist hypothesis testing completely for supposedly superior alternative statistical methods (e.g., Amrhein et al., 2019; Cumming, 2014; Wagenmakers, 2007).

One alternative that has gained notable attention is Bayesian hypothesis testing. In a particularly influential publication in the field of psychology, Rouder et al. (2009) proposed using Bayesian t tests instead of conventional frequentist t tests. Bayesian t tests focus on the relative evidence in the data for one statistical hypothesis, typically referred to as *null hypothesis* ($\mathcal{H}_0$) vis-à-vis another, termed *alternative hypothesis* ($\mathcal{H}_1$). The strength of this evidence is quantified by the *Bayes factor*. As the multiplicative factor for transforming prior beliefs for competing hypotheses to posterior beliefs, it plays a central role in Bayesian hypothesis testing and model comparison (Berger, 2006; Kass & Raftery, 1995). In recent years, the number of applications of Bayesian hypothesis tests in psychological research has been steadily increasing, reflecting their growing importance in the field (Heck et al., in press; Tendeiro & Kiers, 2019).

Martin Schnuerch and Edgar Erdfelder, Department of Psychology, School of Social Sciences, University of Mannheim, Germany. Daniel W. Heck, Department of Psychology, University of Marburg, Germany.

Correspondence concerning this article should be addressed to Martin Schnuerch, B6, 30–32, 68159 Mannheim, Germany. E-mail: martin.schnuerch@psychologie.uni-mannheim.de

Advantages and Limitations of Bayesian Hypothesis Tests

Advocates of Bayesian methods have repeatedly promoted certain properties of Bayes factors as advantageous compared to NHST. For example, as is well known, the p value does not allow analysts to state evidence in favor of the null hypothesis (Gallistel, 2009). Thus, in NHST, we might choose to reject the null but we can never accept it. The Bayes factor, in contrast, can indicate evidence for or against the null, as well as the absence of evidence (Kass & Raftery, 1995; Rouder et al., 2009; Wagenmakers, 2007).

Another limitation of conventional NHST is that *optional stopping*, that is, adaptively increasing the sample size and stopping depending on the data at hand, is inadmissible. If multiple tests are conducted during sampling while stopping only when $p < \alpha$ (where α denotes the predefined Type I error probability) and increasing the sample otherwise, the probability of the p value falling below α at some point approaches one—even when the null hypothesis is true (Armitage et al., 1969). Thus, optional stopping without correction in NHST constitutes a questionable research practice that seriously inflates Type I error rates (John et al., 2012; Simmons et al., 2011). Bayes factors, in contrast, can be computed repeatedly during the sampling process without altering their interpretation as measures of relative statistical evidence (Edwards et al., 1963; Hendriksen et al., 2020; Lindley, 1957; Rouder, 2014). Optional stopping is thus not a problem in Bayesian t tests (but see de Heide & Grünwald, 2020). In fact, sequential Bayes factors have been proposed repeatedly as a means to increase efficiency in hypothesis testing (Rouder, 2014; Schönbrodt et al., 2017; Stefan et al., 2020; Wagenmakers et al., 2012).

Despite their attractive properties, Bayesian t tests also have limitations. The Bayes factor is a continuous evidence measure that indicates how researchers should update their subjective beliefs in competing statistical models or hypotheses. There are currently no normative, theoretically derived thresholds for the Bayes factor that, if exceeded, mandate a decision to reject or accept a hypothesis with controlled error probabilities. Although threshold values for Bayes factors can in principle be calibrated to have the desired error probabilities under the specified statistical models, this calibration may not be straightforward except for atypical applications where both $\mathcal{H}_0$ and $\mathcal{H}_1$ correspond to simple point hypotheses (see Section *Frequentist Error Probabilities in Bayesian t Tests*). In typical applications with more complex specifications of $\mathcal{H}_1$, error probability control for Bayes factors will involve sophisticated calculations and potentially time-consuming simulations (Schönbrodt et al., 2017; Stefan et al., 2020).

In response to the difficulties of error probability control in Bayesian t tests, it has frequently been argued that communicating relative evidence and posterior probabilities (or

odds) directly instead of making dichotomous acceptance/rejection decisions truly reflects the aim of statistical inference in science, that is, what we as researchers actually want to know (e.g., Bakan, 1966; Bayarri et al., 2016; Edwards et al., 1963; Morey et al., 2016; Rozeboom, 1960). However, in contrast to this view, Lakens (2021) argues that it is at least debatable if there really is only one type of inference that all researchers are truly interested in all of the time. In some situations, researchers might indeed be interested in stating the evidence provided by the data or expressing their belief in competing hypotheses after seeing the data. Other contexts, however, compel researchers to decide to act in a certain way, be it by implementing a vaccine for a pandemic disease, by continuing or abandoning a line of research based on the outcome of a pilot study, or simply by making a claim that a hypothesis does or does not hold (Lakens, 2021). In such situations, it is pivotal that the procedure used to arrive at these decisions "in the long run of experience ... not be too often wrong" (Neyman & Pearson, 1933, p. 291). Thus, knowing (and controlling) the properties of a Bayesian t test procedure in terms of how often it results in an erroneous decision in the long rung is highly relevant (Jeon & De Boeck, 2017; Sanborn & Hills, 2014; Sanborn et al., 2014; see also Gelman & Shalizi, 2013). In fact, error probability control in this sense, often referred to as the Neyman-Pearson formulation, is not only important when using Bayesian t tests as a statistical decision-making tool but also from the perspective of statistical inference as *severe testing* of scientific hypotheses or models (Lakatos, 1978; Mayo, 2018; Mayo & Spanos, 2006; Westermann & Hager, 1986).

Aim of This Article

In this article, we discuss a simple extension of Bayesian t tests as a remedy of the aforementioned limitation: We promote a combination of Bayesian t tests with a particular class of frequentist tests, namely, sequential probability ratio tests (SPRTs; Wald, 1947). In SPRTs, hypotheses are tested repeatedly after each observation by calculating a likelihood ratio while increasing the sample size until a predefined threshold is reached. SPRTs for common t-test scenarios have been developed by Rushton (1950, 1952) and Hajnal (1961). Recent simulation studies have demonstrated that they efficiently control the error probabilities of statistical decisions (Schnuerch & Erdfelder, 2020). Importantly, the likelihood ratio underlying these sequential t tests is closely related to the Bayes factor in Bayesian t tests (Stefan et al., 2020). Building on this relationship, we show that Rushton's and Hajnal's sequential t tests can be generalized beyond the case of point hypotheses to a more general class of alternative hypotheses as represented by prior distributions. Consequently, we can use the general logic of the SPRT to derive decision thresholds for sequential Bayes factors such that resulting error probabilities can be controlled explicitly.

We show how this combination of Bayesian t tests and SPRTs unifies the advantages of

the Bayes factor (quantifying statistical evidence) with those of the frequentist test procedure (controlling long-run error rates). To acknowledge Abraham Wald as the originator of the underlying statistical ideas, we call this combination of Bayesian and frequentist inference concepts *Waldian t test.* The proposed procedure draws on previous theoretical advancements in the literature (Berger et al., 1999; Hajnal, 1961; Hendriksen et al., 2020; Wald, 1947) that we apply to three Bayesian t test specifications: the default t test discussed by Rouder et al. (2009), the informed t test introduced by Gronau et al. (2020), and a recently proposed default t test based on so-called *non-local alternative priors* by Pramanik and Johnson (2021).

The remainder of this article is structured as follows: First, we discuss Bayesian t tests and what proper error control in the Neyman-Pearson sense means in this context. We then show how Waldian t tests can achieve this error control. Subsequently, we assess the properties of Waldian t tests in a series of simulations. We also show how to calculate approximate error probabilities of sequential Bayes factor procedures employing conventional Bayes factor thresholds. This is done analytically and thus, does not require time-intensive simulations. Finally, we provide an easy-to-use software tool which implements the proposed Waldian t tests such that psychologists can readily apply them in their own research (https://martinschnuerch.shinyapps.io/Waldian-t-Tests/).

Bayesian t Tests

Inference in Bayesian t tests is based on the Bayes factor, a quantity typically attributed to the works of Sir Harold Jeffreys and Dorothy Wrinch (Wrinch & Jeffreys, 1921). The Bayes factor is the multiplicative factor by which the relative beliefs about two competing hypotheses *before* seeing the data (i.e., the prior odds) are updated in order to arrive at the relative beliefs *after* seeing the data (i.e., the posterior odds; Jeffreys, 1961; Kass & Raftery, 1995):

$$\underbrace{\frac{P(\mathcal{H}_1 \mid \text{data})}{P(\mathcal{H}_0 \mid \text{data})}}_{\text{Posterior odds}} = \underbrace{\frac{p(\text{data} \mid \mathcal{H}_1)}{p(\text{data} \mid \mathcal{H}_0)}}_{\text{Bayes factor}} \cdot \underbrace{\frac{P(\mathcal{H}_1)}{P(\mathcal{H}_0)}}_{\text{Prior odds}} \; . \tag{1}$$

The term $p(\text{data} \mid \mathcal{H}_h)$ denotes the marginal likelihood, subsequently denoted $m_h(\text{data})$, which is the weighted average probability (density) of the observed data under hypothesis h (Rouder et al., 2009). The above formula follows directly from Bayes' rule. It illustrates that the factor by which subjective beliefs should rationally be updated—that is, the extent to which the data should inform what we believe or know—is in fact the relative accuracy of the two hypotheses in predicting the observed data (Rouder & Morey, 2017).

In this article, we focus on the arguably most common scenario in statistical inference,

namely, the test of mean differences between two independent samples with common but unknown standard deviation σ. Let x_i and y_j denote observations from each group, with $i = 1, \ldots, n_x$ and $j = 1, \ldots, n_y$, modeled as

$$
\begin{aligned}
X_i &\sim \mathrm{Normal}(\mu + \tfrac{\delta\sigma}{2}, \sigma^2), \\
Y_j &\sim \mathrm{Normal}(\mu - \tfrac{\delta\sigma}{2}, \sigma^2).
\end{aligned}
\tag{2}
$$

In this notation, the unknown grand mean μ and population standard deviation σ are so-called nuisance parameters while δ denotes the test-relevant parameter of primary interest, namely, the standardized population effect size (i.e., Cohen's d; Cohen, 1988). The hypotheses can thus be expressed in terms of δ. In the classical setting that also underlies conventional NHST t tests, the null hypothesis posits that there is no difference between the two population means, that is, the effect is absent, $\mathcal{H}_0 : \ \delta = 0$. Correspondingly, the two-sided alternative hypothesis specifies that the means differ, that is, there is some non-zero effect, $\mathcal{H}_1 : \ \delta \neq 0$.

A Bayesian test of these hypotheses requires the specification of prior distributions on unknown model parameters. A particularly influential development in the choice of priors is due to Jeffreys (1961). In this article, we consider three extensions of Jeffreys's t test, namely those discussed by Rouder et al. (2009), Gronau et al. (2020), (both of which are implemented in the open-source statistics software JASP; JASP Team, 2020), and Pramanik and Johnson (2021). In the following, we adopt Gronau et al.'s notation to outline the general framework that encompasses all three Bayesian t tests. Subsequently, we discuss the specific prior settings underlying each of the three approaches. Readers interested in the mathematical details and derivations are referred to the original articles.

A General Framework

Under the null hypothesis, the specified statistical model (Equation 2) has two free parameters, μ and σ. Under the alternative hypothesis, δ is an additional free parameter. Let $\pi(\mu, \sigma \mid \mathcal{H}_0)$ and $\pi(\delta, \mu, \sigma \mid \mathcal{H}_1)$ denote the prior distributions of the statistical models corresponding to the null and the alternative hypothesis, respectively. The marginal likelihood of each hypothesis is calculated by integrating the conditional probability (density) of the observed data across all possible parameter values weighted by the respective prior distributions. Consequently, the Bayes factor for the Bayesian t test is given by

$$
\mathrm{BF}_{10} = \frac{m_1(\mathbf{x}, \mathbf{y})}{m_0(\mathbf{x}, \mathbf{y})} = \frac{\int \int \int p(\mathbf{x}, \mathbf{y} \mid \delta, \mu, \sigma, \mathcal{H}_1) \pi(\delta, \mu, \sigma \mid \mathcal{H}_1) \, \mathrm{d}\delta \, \mathrm{d}\mu \, \mathrm{d}\sigma}{\int \int p(\mathbf{x}, \mathbf{y} \mid \mu, \sigma, \mathcal{H}_0) \pi(\mu, \sigma \mid \mathcal{H}_0) \, \mathrm{d}\mu \, \mathrm{d}\sigma},
\tag{3}
$$

where $\mathbf{x} = (x_1, \ldots, x_{n_x})$ and $\mathbf{y} = (y_1, \ldots, y_{n_y})$ denote the observed data from the two groups and $p(.)$ is the probability (density) of the data conditional on specific values of the parameters under each hypothesis.

Equation 3 highlights the relevance of the choice of prior distributions when calculating and interpreting a Bayesian t test. The prior distributions define the statistical models compared by means of the Bayes factor as statistical representations of the underlying (substantive) hypotheses (Vanpaemel, 2010). These statistical models are not identical to the hypotheses, however. Consider the alternative hypothesis that $\delta \neq 0$. This is a *composite hypothesis* because without a restriction of δ to a specific value it does not admit a particular probability distribution of the data but rather a complete family of probability distributions. By specifying and marginalizing across a prior distribution, however, the statistical model provides a single, specific probability distribution of the data, namely, the marginal likelihood. Thus, the statistical hypothesis actually tested becomes a *simple hypothesis* without free parameters. This hypothesis states that the probability distribution of the observed data $\mathbf{x}, \mathbf{y}$ is given by the marginal likelihood function (Berger et al., 1997, 1999):

$$\mathcal{H}_1: \ \mathbf{x}, \mathbf{y} \sim m_1(\mathbf{x}, \mathbf{y}). \tag{4}$$

We may write, accordingly,

$$\mathcal{H}_0: \ \mathbf{x}, \mathbf{y} \sim m_0(\mathbf{x}, \mathbf{y}), \tag{5}$$

stating that the probability distribution of the observed data is given by the marginal likelihood of the null hypothesis.

How to choose the priors that define the statistical models at test? A common choice for the nuisance parameters μ and σ is the so-called right Haar prior $\pi(\mu, \sigma) \propto 1/\sigma$. The choice is uncritical as it is a non-informative prior and identical for both statistical models $\mathcal{H}_0$ and $\mathcal{H}_1$ (Ly et al., 2016). Moreover, when using this prior, the Bayes factor can be expressed as a function of the observed outcome of a classical t test, $t = \sqrt{n}(\bar{X} - \bar{Y})/\hat{\sigma}_p$, where $n = n_x n_y/(n_x + n_y)$ denotes the effective sample size and $\hat{\sigma}_p = \sqrt{(n_x - 1)s_x^2 + (n_y^2 - 1)s_y^2} \cdot \nu^{-\frac{1}{2}}$ denotes the pooled standard deviation with $\nu = n_x + n_y - 2$ degrees of freedom. Let $\mathrm{T}_\nu(t \mid \Delta)$ be the density of a t distribution with ν degrees of freedom and non-centrality parameter $\Delta = \sqrt{n}\delta$, then the Bayes factor in Equation 3 can be expressed as

$$\mathrm{BF}_{10} = \frac{\int \mathrm{T}_\nu(t \mid \sqrt{n}\delta)\pi(\delta \mid \mathcal{H}_1) \, \mathrm{d}\delta}{\mathrm{T}_\nu(t)}. \tag{6}$$

For any proper prior $\pi(\delta \mid \mathcal{H}_1)$, the Bayes factor can be calculated by simple numerical

integration across all possible δs.[1] The Bayesian t tests considered herein differ only with respect to this prior, that is, the hypothesized distribution of δ under the alternative hypothesis.

Prior t Distributions

As a general framework for a prior on δ, Gronau et al. (2020) propose a scaled t distribution,

$$\pi(\delta \mid \mathcal{H}_1) = \frac{1}{\gamma} \mathrm{T}_\kappa \left(\frac{\delta - \mu_\delta}{\gamma} \right),$$

where the scale parameter γ, the degrees-of-freedom parameter κ, and the location parameter μ_δ are defined by the analyst before the analysis. Thereby, precise expectations about how large and how variable the effect is can be incorporated into the test (see Figure 1). We will refer to tests based on such prior specifications with $\mu \neq 0$ as *informed t* tests.

The informed Bayesian t test is particularly useful when researchers test a theoretically motivated non-zero effect. If, based on the theory at test, a certain effect is expected, that is, a certain deviation from $\delta = 0$, this expectation should be represented by the statistical model by defining the location parameter μ_δ accordingly. By specifying γ and κ, the flexible t distribution also allows for the quantification of uncertainty about the assumed effect size or the specification of a random effect, that is, true variation in the exact value of the effect across experiments.

Another advantage of Gronau et al.'s prior is that it encompasses the specification discussed by Rouder et al. (2009) as a special case. Setting $\mu_\delta = 0$ and $\kappa = 1$, we obtain the central Cauchy distribution proposed as a default prior under the alternative hypothesis for Bayesian t tests. The justification for the Cauchy is primarily based on its favorable mathematical properties (for details, see Ly et al., 2016; Rouder et al., 2009). We also see a reasonable substantive interpretation and practical appeal, however: Given a theoretically motivated null hypothesis that two group means are expected to be equal—that is, a theoretically motivated invariance—there might be no reason to assume a certain fixed effect size or even a direction under the alternative. For such a case, the Cauchy may be interpreted as a representation of the distribution of possible non-zero effect sizes in the field when the hypothesized invariance does not hold. It represents true effects randomly distributed around the effect predicted by the null hypothesis. Reasonably, most of these effects can be expected to be small effects close to the null, where the Cauchy distribution puts most of its weight. At the same time, there is also substantial emphasis on large effect

---

[1]In some cases (e.g., in the context of the non-local alternative priors proposed by Pramanik & Johnson, 2021), this integral can be solved analytically.

sizes that are less prevalent but not impossible. In the following, we will denote a Bayesian t test based on a central Cauchy prior under the alternative hypothesis as *default t* test.

Non-Local Alternative Priors

Despite the appeal of the Cauchy distribution as a default prior on δ under the alternative, Johnson and Rossell (2010) note that, paradoxically, the Cauchy assigns the highest probability density to the value consistent with the null hypothesis (i.e., $\delta = 0$). Not only may this be unsettling from a substantive point of view, it also comes with technical difficulties (but see van Ravenzwaaij & Wagenmakers, in press): Outcomes consistent with the null hypothesis are assigned a high probability also under the alternative hypothesis. Consequently, the rate with which the Bayes factor accumulates evidence for the null and the alternative hypothesis is highly asymmetrical: Whereas strong evidence *against* the null hypothesis accumulates rapidly for data inconsistent with the hypothesis as the sample size increases, outcomes consistent with the null hypothesis require much larger sample sizes to indicate compelling evidence *in favor* of it.

As a remedy, Pramanik and Johnson (2021) recently proposed a different default prior under the alternative hypothesis when testing a substantively motivated null hypothesis. Unlike the Cauchy distribution, the proposed non-local alternative prior (NAP) assigns a probability density of 0 to the value of δ that is consistent with the null hypothesis. Thus, compared to the Bayesian t test based on the Cauchy prior, a Bayes factor based on an NAP may be more acceptable from a substantive point of view, and has more symmetrical (and potentially higher) evidence accumulation rates. Pramanik and Johnson (2021) suggest a normal moment prior on δ, that is,

$$\pi(\delta \mid \mathcal{H}_1) = \frac{(\delta - \mu_\delta)^2}{\tau^2 \sigma^2} \, \phi(\delta \mid \mu_\delta, \tau^2 \sigma^2),$$

where $\phi(q \mid a, b)$ denotes the density of a normal distribution with mean a and variance b evaluated at q. To make this prior a *non-local* prior, that is, put a prior density of 0 on the value specified under the null hypothesis, we set $\mu_\delta = \delta_0 = 0$. Furthermore, because the prior is specified for a standardized effect size, we let $\sigma^2 = 1$, leaving τ^2 as the only free parameter. As a default choice for hypothesis tests in behavioral research where most non-zero effects can be expected to be of small to medium size, Pramanik and Johnson (2021) recommend setting $\tau^2 = 0.045$. Although this prior represents a proposed default under the alternative, we refer to a Bayesian t test with an NAP prior as *NAP t* test to distinguish it from Rouder et al.'s default t test based on the Cauchy prior.

Figure 1 illustrates the prior distributions on the effect size in the three Bayesian t
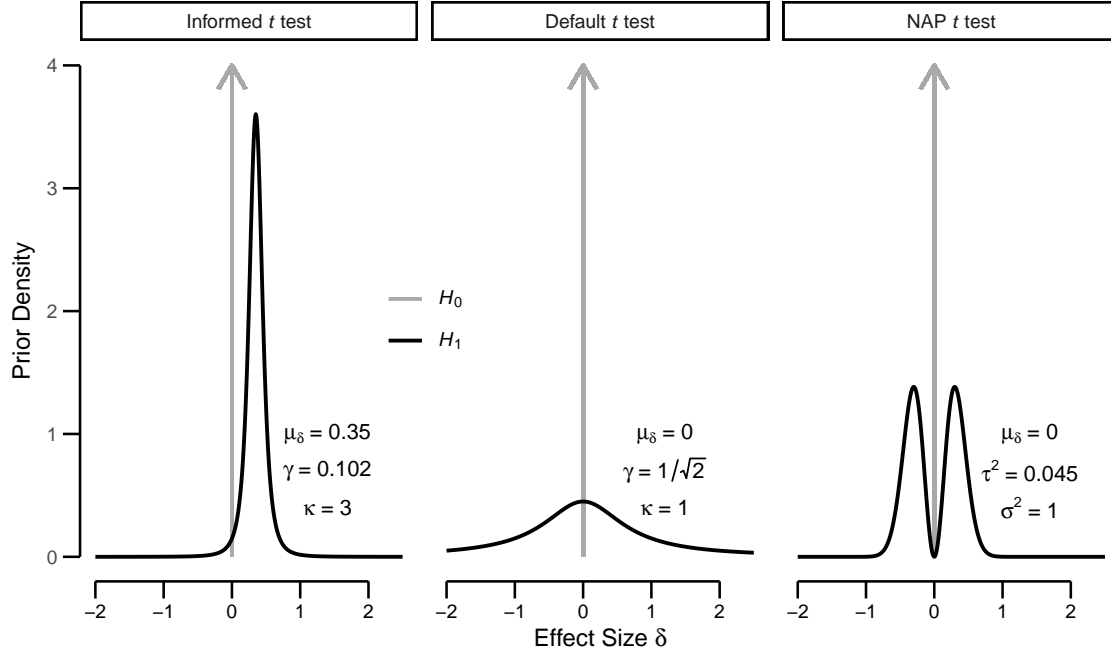
Figure 1. Illustration of prior distributions in Bayesian t tests. Prior under the null hypothesis: $\delta = 0$; Prior under the alternative hypothesis in the default and informed t tests: scaled t distribution with κ degrees of freedom, location μ_δ, and scale γ; Prior under the alternative hypothesis in NAP t tests: normal moment prior with location μ_δ and variance $\tau^2 \sigma^2$

tests presented above. While the null hypothesis is represented by a point prior on $\delta = 0$ in all specifications, the informed test places a t distribution (in this example with $\kappa = 3$ degrees of freedom, $\gamma = 0.102$, and $\mu_\delta = 0.350$) on δ under the alternative hypothesis. This particular prior is based on an example presented by Gronau et al. (2020). The default test uses a scaled Cauchy distribution instead, which is obtained by specifying $\kappa = 1$ and $\mu_\delta = 0$. Moreover, the prior scale $\gamma = 1/\sqrt{2}$ is often used as a default setting in software implementations, for instance, in the *BayesFactor* package in R (Morey & Rouder, 2015) or in JASP. The NAP test, in contrast, places a normal moment prior as recommended by Pramanik and Johnson (2021) and specified above on δ. Note that both the default and the NAP t test clearly correspond to two-sided alternative hypotheses, as they are symmetric and place half of their weight on either side of $\delta = 0$. The informed t test with the given specification, in contrast, rather corresponds to a one-sided test as it places only around 2% of its weight on effect sizes smaller than 0. In principle, all prior distributions could be truncated at 0 to test one-sided hypotheses. Without loss of generality, however, we focus on non-truncated prior distributions in this article. While the informed test is appropriate when the theoretically motivated hypothesis predicts some non-zero effect, the default and

the NAP test are more appropriate when testing a substantively motivated invariance.

Error Probabilities

From a Bayesian perspective, inference should be about inductive probabilistic statements about hypotheses conditional on the observed data (Jeffreys, 1961). Importantly, this does not require any reference to procedural properties such as long-run error rates given that one of the hypotheses is assumed to be true (Wagenmakers & Gronau, 2018). Rouder (2014) summarizes that "the key to understanding Bayesian analysis is to focus on the degree of belief for considered models, which need not and should not be calibrated relative to some hypothetical truth" (p. 308).

The view that Bayesian inference is purely inductive has been challenged, however (e.g., Gelman & Shalizi, 2013). As we argue above, we believe that different contexts impose different desiderata on any procedure of statistical inference. While methods for quantifying statistical evidence and updating subjective beliefs certainly have their place in scientific research, they constitute but one tool in a researcher's toolkit. Therefore, we endorse the view that consideration and control of frequentist properties is an important endeavor also in the context of Bayesian analysis (see Berger & Bayarri, 2004; Gu et al., 2016; Jeon & De Boeck, 2017; Sanborn et al., 2014).

The focus of this article is on error probabilities. We may differentiate between two kinds of error probabilities (Wagenmakers & Gronau, 2018), however, and this distinction is particularly important in the interplay of Bayesian and frequentist procedures. We call the first kind *Bayesian error probabilities*: A Bayesian analysis of competing hypotheses renders posterior probabilities, that is, probabilities of the hypotheses conditional on the observed data (see Equation 1). Imagine a researcher who obtains $P(\mathcal{H}_1 \mid \text{data}) = .90$. Given this result, what is the probability that the researcher errs on rejecting the null hypothesis? The answer is simple from a Bayesian perspective: It is the extent to which the researcher believes in the null hypothesis after seeing the data, that is, the posterior probability $P(\mathcal{H}_0 \mid \text{data}) = 1 - P(\mathcal{H}_1 \mid \text{data}) = .10$. The important characteristic of Bayesian error probabilities is that they are conditional both on the observed data and the specified prior probabilities of the hypotheses. Let α_B and β_B denote the Bayesian (conditional) probabilities of falsely rejecting the null and the alternative hypothesis, respectively. For equal prior probabilities, that is, $P(\mathcal{H}_h) = .50$ for $h = 0, 1$, the Bayesian error probabilities are given by

$$\alpha_\mathrm{B} = \frac{1}{\mathrm{BF}_{10} + 1},$$
$$\beta_\mathrm{B} = \frac{\mathrm{BF}_{10}}{\mathrm{BF}_{10} + 1}. \tag{7}$$

Equation 7 illustrates an interesting property of Bayesian error probabilities: If the Bayes factor BF_{10} tends to infinity (i.e., if we gain extreme evidence in favor of the alternative hypothesis), then α_B goes to 0. Thus, when conditioned on data that provide infinite support for the alternative hypothesis, the conditional probability to commit an error by rejecting the null is zero. The same holds true for β_B if the data show unequivocal evidence for the null hypothesis (i.e., if $BF_{10} \to 0$). In this case, the Bayesian probability to err on accepting the null is zero.

The second kind of error probabilities is the one we call *frequentist error probabilities*. The critical difference to the Bayesian kind is that these error probabilities do not reflect subjective belief or uncertainty and are thus not conditional on a particular set of observed data or prior probabilities of hypotheses. Rather, they are stable properties of the inference procedure under the assumption that either $\mathcal{H}_0$ or $\mathcal{H}_1$ is true. The rationale underlying these error probabilities is that they characterize the accuracy of an inference procedure in the long run. This focus is intimately linked to a behaviorist interpretation of hypothesis testing as a rule to guide decisions: We either reject the null hypothesis or we accept it, meaning that without any reference to whether we believe the hypotheses to be true or false we take a specific course of action with regard to them (Lakatos, 1978; Neyman & Pearson, 1933). By doing so, we commit a Type I error if we reject the null hypothesis when it is true, or a Type II error if we accept the null when it is false. To ensure the quality of our decisions, we require that the procedure (or rule) on which the decisions are based have low probabilities to commit either error (Neyman, 1977).

Formally, the frequentist error probabilities of a statistical hypothesis test are defined as

$$\begin{aligned}
\alpha &= P(\text{reject } \mathcal{H}_0;\ \mathcal{H}_0), \\
\beta &= P(\text{reject } \mathcal{H}_1;\ \mathcal{H}_1),
\end{aligned} \tag{8}$$

where $P(\text{reject } \mathcal{H}_h;\ \mathcal{H}_h)$ denotes the probability of the test to reject hypothesis h ($h = 0, 1$) when it is true. That is, for data generated under the statistical models specified in the test, the probabilities to reject the corresponding hypotheses are α and β, respectively. The advantage of Neyman-Pearson tests (and Wald's sequential tests, as we outline below) is that for given statistical hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$, the procedure can be designed to satisfy certain error probabilities. Thus, by choosing an appropriate experimental design (e.g., based on an a priori power analysis; Cohen, 1988) researchers can ensure that the test of their statistical hypotheses has a sufficiently small risk of an erroneous decision.

It has been argued that frequentist error probabilities are only relevant or appropriate in pre-experimental considerations but no longer when data have been observed. According to this argument, they only reflect properties of the procedure and ignore the specific

information provided by the observed data (Berger et al., 1997; Wagenmakers & Gronau, 2018). This criticism has been challenged, however, by the severity approach introduced and most prominently represented by the philosopher Deborah Mayo as an improvement of Popper's well-known *critical rationalism* (e.g., Mayo, 1996, 2018; Mayo & Spanos, 2006; see also Lakatos, 1978; Westermann & Hager, 1986). According to this perspective, frequentist error probabilities do not only provide insight into long-run properties that are relevant for planning a study or for the totality of performed tests. Instead, frequentist error probabilities also provide a means to assess the severity with which a single substantive hypothesis has been tested. A test of a certain hypothesis is more severe (i.e., more rigorous) if it has a higher probability to detect a deviation from what we expect under the hypothesis. Although low frequentist error probabilities do not provide a direct measure of this severity, they are a necessary condition for a severe test (Lakatos, 1978; Mayo & Spanos, 2006; Westermann & Hager, 1986). Therefore, not only from a behaviorist, decision-making perspective but also from a philosophy-of-science point of view, frequentist error probabilities are relevant and worth considering or, better yet, controlling (Sanborn et al., 2014).

Frequentist Error Probabilities in Bayesian t Tests

The plea to consider frequentist concepts such as error probabilities and statistical power ($1-\beta$, the complement of the Type II error probability) also in the context of Bayesian t tests is not particularly new. For example, James Berger has long argued that "statisticians should readily use both Bayesian and frequentist ideas" (Berger & Bayarri, 2004, p. 58). There have been previous efforts in the literature to *estimate* error probabilities of Bayesian t tests by simulation (e.g., Jeon & De Boeck, 2017; Sanborn & Hills, 2014; Schnuerch & Erdfelder, 2020; Schönbrodt et al., 2017; Yu et al., 2014), to *control* them explicitly (e.g., Gu et al., 2016; Hoijtink et al., 2016; Schönbrodt & Wagenmakers, 2018; Stefan et al., 2020), and to unify Bayesian and frequentist test procedures (e.g., Bayarri et al., 2016; Berger, 2003; Berger et al., 1997, 1999). In contrast to the frequentist error probabilities that we consider in this article, however, such unification efforts have focused mostly on error rates conditional on the observed data (Berger et al., 1994).

The key to controlling unconditional frequentist error probabilities is the consideration of a reference set of possible outcomes of a specific test procedure. Neyman-Pearson tests are based on possible outcomes of a hypothesis test with a fixed sample size. If the sampling distribution of the test statistic is known under $\mathcal{H}_0$ and $\mathcal{H}_1$, error probabilities are controlled by choosing an appropriate critical value of the test statistic and a minimal sample size. In the same vein, if the sampling distribution of the Bayes factor is known, it is straightforward to determine an acceptance/rejection threshold such that the probability for the Bayes factor to exceed this threshold under certain population scenarios is controlled. Consider the

following example: We wish to perform a Bayesian t test with controlled error probabilities under the two point hypotheses $\mathcal{H}_0\colon \delta = \delta_0$ and $\mathcal{H}_1\colon \delta = \delta_1$. For a given sample size N, the sampling distribution of the test statistic T_N is known under both hypotheses. Let t_{crit} denote the critical value of the test statistic such that $P(T_N > t_{\mathrm{crit}};\ \delta_0) = \alpha$. If the Bayes factor can be expressed as a strictly monotonically increasing function of the test statistic, $\mathrm{BF}(T_N)$, it follows that $P(\mathrm{BF}(T_N) > \mathrm{BF}(t_{\mathrm{crit}});\ \delta_0) = P(T_N > t_{\mathrm{crit}};\ \delta_0) = \alpha$. Equivalently, $P(\mathrm{BF}(T_N) > \mathrm{BF}(t_{\mathrm{crit}});\ \delta_1) = P(T_N > t_{\mathrm{crit}};\ \delta_1) = 1 - \beta$. Thus, a Bayesian t test with a decision threshold $\mathrm{BF}(t_{\mathrm{crit}})$ has frequentist error probabilities α and β under $\delta = \delta_0$ and δ_1, respectively.

If, however, the sampling distribution of the Bayes factor is unknown, it is not possible to analytically determine threshold values to control frequentist error probabilities. In a Bayesian t test, the alternative hypothesis as represented by the prior does not correspond to a point, as in the above example, but to a distribution. To our knowledge, controlling the frequentist error probabilities under such a statistical model in a fixed-sample design would require a numerical or a simulation approach. Such a method is implemented in *Bayes factor design analysis*, where threshold values for the Bayes factor and required sample sizes are calibrated by means of extensive Monte Carlo simulations (Schönbrodt & Wagenmakers, 2018; Stefan et al., 2019; Stefan et al., 2020).

In our view, the above mentioned efforts to assess and control frequentist properties in the context of Bayesian hypothesis tests have made valuable contributions to the statistics toolkit. To complement them, we propose a procedure that (i) aims at unifying Bayesian t tests with frequentist error probability control in the Neyman-Pearson sense, (ii) retains a fully Bayesian interpretation, and (iii) does not require any simulations.

In contrast to fixed-sample designs, our analytical approach considers a reference set of possible outcomes that provide the same degree of evidence instead of having the same sample size. This is the central idea of what is known as *sequential analysis* (Barnard, 1949; Wald, 1947). Considering a reference set of experimental outcomes with a certain (minimum) amount of statistical evidence provides a strikingly simple and efficient means to control error probabilities under the specified statistical hypotheses. This idea is at the heart of the PRT which was developed and introduced by Abraham Wald in the 1940s (Wald, 1947) and which also provides the basis for the Waldian t test we propose herein.

Sequential Probability Ratio Tests

In contrast to conventional statistical tests, sequential procedures dispense with the requirement to define a fixed sample size a priori. Instead, the data are sampled sequentially: After any new step of the sampling process (potentially after every single observation), the

test requires a decision to either continue or terminate sampling. This decision is based on a stopping rule, and the choice of this rule defines the long-run properties of the sequential procedure (i.e., expected sample size and error rates). On average, sequential tests require substantially smaller sample sizes than conventional fixed-sample tests. This renders them notably more efficient for many situations (Lakens, 2014; Lang, 2017; Schnuerch & Erdfelder, 2020; Schönbrodt et al., 2017).

The stopping rule employed in the SPRT is based on the likelihood ratio (LR). The LR denotes the probability (density) of the data under one set of specific parameter values relative to that under a different set. As before, let $\mathbf{x} = (x_1, \ldots, x_{n_x})$ and $\mathbf{y} = (y_1, \ldots, y_{n_y})$ be the observed data, and $p(\mathbf{x}, \mathbf{y}; \theta)$ the probability function of the data given the true parameter vector $\theta = (\delta, \mu, \sigma)$. We are interested in testing the simple hypothesis $\mathcal{H}_0 \colon \theta = \theta_0$ against $\mathcal{H}_1 \colon \theta = \theta_1$. Similar to Equations 4 and 5, these hypotheses can be expressed in terms of the probability distribution that they imply, that is,

$$
\begin{aligned}
\mathcal{H}_0 &\colon \mathbf{x}, \mathbf{y} \sim p(\mathbf{x}, \mathbf{y}; \theta_0), \\
\mathcal{H}_1 &\colon \mathbf{x}, \mathbf{y} \sim p(\mathbf{x}, \mathbf{y}; \theta_1).
\end{aligned}
\tag{9}
$$

For the sequential test of these hypotheses, the likelihood ratio is computed after each additional observation, starting at $n_x = n_y = 1$:

$$
\text{LR}_N = \frac{p(\mathbf{x}, \mathbf{y}; \theta_1)}{p(\mathbf{x}, \mathbf{y}; \theta_0)},
\tag{10}
$$

and at each step, one of the following three decisions is made:

1) Accept $\mathcal{H}_1$ and reject $\mathcal{H}_0$ when $LR_N \geq A$;
2) Accept $\mathcal{H}_0$ and reject $\mathcal{H}_1$ when $LR_N \leq B$; $\hfill$ (11)
3) Sample a new independent observation x_{n_x+1} or y_{n_y+1} when $B < LR_N < A$.

To complete the specification of the SPRT's stopping rule, appropriate values for the decision boundaries A and B must be defined. Wald (1947) showed that the determination of these boundaries in practical applications, such that the procedure does not exceed certain error probabilities α and β, is straightforward. By definition of Rule 1 in Equation 11, any sample that leads the analyst to accept $\mathcal{H}_1$ satisfies the following inequality,

$$
p(\mathbf{x}, \mathbf{y}; \ \theta_1) \geq A \cdot p(\mathbf{x}, \mathbf{y}; \ \theta_0),
$$

indicating that this sample is at least A times more likely to occur under $\mathcal{H}_1$ than under $\mathcal{H}_0$.

This means that the probability to obtain any sample that leads to the acceptance of $\mathcal{H}_1$ is at least A times larger under $\mathcal{H}_1$ than under $\mathcal{H}_0$. The probability to obtain such a sample, in turn, is equivalent to the total probability of accepting $\mathcal{H}_1$. Thus, the probability to accept the alternative hypothesis with the given procedure is at least A times larger under the alternative than under the null. In our notation (see Equation 8), the latter is defined as the Type I probability α and, because the SPRT eventually terminates with accepting either of the specified hypotheses (see Wald, 1947, Appendix A.1, for a proof), the former is defined as $1 - \beta$ (statistical power). Hence, $1 - \beta \geq A\alpha$. Following the same reasoning for a sample that leads the researcher to accept $\mathcal{H}_0$, we obtain $\beta \leq B(1 - \alpha)$. Rewriting these inequalities shows that $A \leq (1 - \beta)/\alpha$ and $B \geq \beta/(1 - \alpha)$.

Wald (1947) proved that treating these inequalities as equalities to define the boundary values of the SPRT in practice "cannot result in any appreciable increase in the value of either α or β" (p. 46). In fact, because LR_N almost always strictly exceeds the boundary at termination (a phenomenon called *overshooting*), the resulting error rates of the procedure will undercut the nominal α and β. If exact error control is required, the threshold values A and B can be numerically adjusted to result in the nominal error rates for the specified hypotheses (Stefan et al., 2020). However, Wald (1947) conjectured that the decrease in the SPRT's efficiency due to its conservative behavior is negligible, and recent simulation studies have corroborated this supposition (Schnuerch & Erdfelder, 2020; Schnuerch et al., 2020). Moreover, the availability of simple analytical solutions increases the likelihood that the approach is understood and used by substantive researchers. Thus, to construct an SPRT for the test of two simple hypotheses with upper-bound error probabilities α and β, a researcher may follow the sampling plan given by Equation 11 with threshold values defined by

$$
\begin{aligned}
A &= \frac{1 - \beta}{\alpha}, \\
B &= \frac{\beta}{1 - \alpha}.
\end{aligned}
\tag{12}
$$

Sequential Bayesian t Tests

Sequential sampling plans have also been proposed for Bayesian t tests (Schönbrodt et al., 2017; Wagenmakers et al., 2012). The procedure is typically referred to as sequential Bayes factors (SBFs). Similar to the likelihood ratio in the SPRT, the Bayes factor in SBFs is calculated repeatedly during the sampling process until a decision is made to terminate the procedure. According to the likelihood principle (Berger & Wolpert, 1988), the interpretation of the Bayes factor as a measure of evidence does not depend on how the data were sampled. Consequently, this interpretation remains unaltered also in sequential applications, thus giving justification to sequential Bayesian t tests (Hendriksen et al., 2020; Rouder, 2014; but

see de Heide & Grünwald, 2020).

Stefan et al. (2020) recently pointed out that SBFs and SPRTs can be regarded as examples of a general sequential testing framework. In this framework, appropriate threshold values must be chosen to control error probabilities. For SBFs, these threshold values can be found by simulating the procedure for a range of possible thresholds with data generated from a specific effect size (i.e., based on Bayes factor design analysis). We agree with Stefan et al. that, apart from fundamental differences in philosophical foundations, SBFs and SPRTs can be seen as two instances of the same conceptual framework. Both are based on sequentially computed likelihood ratios with predefined thresholds, one on simple likelihood ratios (SPRTs) and the other on marginal likelihood ratios (SBFs). In contrast to Stefan et al., however, we maintain that error probability control does not require a simulation-based approach to SBFs assuming specific effects under $\mathcal{H}_1$. Rather, it can be conducted analytically based on the general theory of the SPRT assuming either specific effects as in Hajnal's t test (see Schnuerch & Erdfelder, 2020) or an effect size distribution under $\mathcal{H}_1$. This analytic approach is more efficient in practice and also promotes a deeper understanding of the substantive meaning of the statistical models at test, as well as the theoretical links and underlying assumptions of SBFs and SPRTs.

We fully exploit the consequences of the SBF-SPRT relationship in this article. We argue that sequential Bayesian t tests can be construed as a special case of the SPRT. This unified procedure, which we call Waldian t test, controls frequentist error probabilities as defined in Equation 8 for the statistical models specified by the prior distributions. The general idea of such a procedure was outlined already in 1947 by Abraham Wald and also briefly discussed by Berger et al. (1999). Further mathematical justifications can be found in Hendriksen et al. (2020).

Waldian t Tests

Waldian t tests combine Bayesian t tests with an SPRT's sampling plan. They are defined by the stopping rules given in Equation 11 where the test statistic LR_N is replaced by $\mathrm{BF}_{10}(N)$, that is, the Bayes factor given in Equation 6, where N denotes the given sample size across both groups. Further, the threshold values A and B are defined according to the simple formulas suggested by Wald (1947) and given in Equation 12. Consequently, a Waldian t test is carried out by continuing to sample independent observations x_i and y_j as long as

$$\frac{\beta}{1 - \alpha} < \mathrm{BF}_{10}(N) < \frac{1 - \beta}{\alpha}, \tag{13}$$

and terminating and accepting $\mathcal{H}_0$ or $\mathcal{H}_1$ as soon as the left or right inequality is violated, respectively. The resulting procedure has frequentist error probabilities as defined in Equation

8, with α and β serving as upper bounds under the statistical null and alternative hypothesis, respectively.

Priors Imply Simple Hypotheses

One might object that replacing LR_n by $\text{BF}_{10}(N)$ is not justified because the former is based on simple hypotheses and the latter on composite hypotheses. It is true that the SPRT is generally based on the specification of two competing point hypotheses (e.g., $\delta = \delta_0$ versus $\delta = \delta_1$). The Bayes factor, in contrast, specifies prior distributions for the unknown parameters (e.g., $\delta \sim \pi(\delta \mid \mathcal{H}_1)$). The distinction between a point and a distributional hypothesis is not identical to that between a simple and a composite hypothesis, however. As we explained above, by specifying a proper prior distribution and marginalizing over it, the probability distribution of the observed data is completely specified under each hypothesis and does not depend on any free parameters anymore. Thus, the two *statistical* hypotheses actually compared by the Bayes factor are always simple hypotheses as defined by the specified priors (Berger et al., 1997, 1999; see also Wald, 1947).

We introduced $m_h(\text{data})$ as the marginal likelihood of hypothesis h ($h = 0, 1$). Conceptually, a point hypothesis is just a special case of a prior distribution on a parameter. Thus, we may use the universal notation introduced in Equations 4 and 5 based on marginal likelihoods to represent the likelihood ratio as well as the Bayes factor by the ratio $m_1(\text{data})/m_0(\text{data})$, where $m_h(\text{data})$ denotes the likelihood of hypothesis h marginalized across either a continuous prior distribution, that is, $m_h(\text{data}) = \int_\Theta p(\text{data} \mid \theta)\,\pi(\theta \mid \mathcal{H}_h)\,\mathrm{d}\theta$, or a prior point mass, that is, $m_h(\text{data}) = p(\text{data} \mid \theta_h)$.

Frequentist Justification

What are the implications of using a marginal likelihood ratio for the SPRT? Let us first consider the prior on the nuisance parameters (μ and σ). Wald (1947) himself suggested to use a prior (although he used the term *weight function*) very similar to the right Haar prior for the unknown scale parameter to construct a sequential t test. In fact, the denominator of the Bayesian t test as presented in Equation 6 is identical to that in the likelihood ratio of Rushton's and Hajnal's sequential probability ratio t tests (Schnuerch & Erdfelder, 2020). de Heide and Grünwald (2020) refer to the right Haar prior as "Type 0 prior" because it has favorable properties in the context of optional stopping: As Hendriksen et al. (2020) showed, when using the right Haar prior in the Bayesian t test, the Type I error probability has a fixed value for any true (μ, σ) in the admissible parameter space. Thus, if $\delta = 0$, a Waldian t test as defined in Equation 13 has the desired Type I error probability α irrespective of the true values of the nuisance parameters.

This result also holds true for the error probability under the alternative hypothesis. In this case, however, δ is no longer constrained to a single point in the parameter space. Hence, the Type II error probability β is a single-valued function of the parameter δ in the parameter space, that is, $\beta(\delta)$ for $\delta \in \Omega_1$. Therefore, for a Waldian t test as defined in Equation 13, the error probability under the alternative hypothesis does not refer to any specific effect size. Instead, it is an expected error probability, that is, a weighted average of error probabilities across Ω_1:

$$\int_{\delta \in \Omega_1} \beta(\delta) \; \pi(\delta \mid \mathcal{H}_1) \; \mathrm{d}\delta = \beta.$$

Hendriksen et al. (2020, p. 8) refer to this result as *semi-frequentist* because controlling the expected error probability might appear unsatisfactory from a frequentist point of view. We believe, however, that this procedure has a fully frequentist justification in many situations, and that in these situations, the average error probability constitutes a proper frequentist error probability as defined in Equation 8 (see also Gillett, 1994). As we noted above, a prior distribution can be attached with appealing frequentist interpretations, for example, as a model of the variation of true effect sizes across different experiments. This random-effect notion is routinely employed in meta-analysis and hierarchical modeling, for example (Borenstein et al., 2009; Hedges & Olkin, 1985). It reflects the reasonable position that the true effect size, even if the studied underlying mechanism is the same, is influenced by characteristics of the experiment and the participants. Thus, variation in true effects across studies is to be expected (Ulrich et al., 2018), and the prior distribution represents the analyst's hypothesis about this variation. Consequently, proper error control is achieved if we control the expected error probability given the distribution of effect sizes across studies, not the error probability associated with a specific effect size selected more or less arbitrarily from this distribution.

Another appealing interpretation is that of a "power weight function" (Bayarri et al., 2016, p. 101). According to this interpretation, the prior defines certain regions of the admissible range of effect sizes for which high statistical power is desired, while controlling the expected error probability (and thus, the expected statistical power) across the entire range. In sum, we argue that a Waldian t test has a fully frequentist justification whenever there is a meaningful substantive interpretation of $\pi(\delta \mid \mathcal{H}_1)$ with respect to properties of δ.

Bayesian Justification

A Bayesian may hold different standards as to what makes the interpretation of a prior meaningful. From a subjective perspective, priors and the statistical models they

define are mathematical abstractions of substantive hypotheses, not reflections of existing data-generating processes (Rouder et al., 2016). Accordingly, the interpretation of the Bayes factor as statistical evidence for the specified models holds, irrespective of how the data were sampled and what the true data generating process may be. This position is why Bayesians have stressed that sequential testing (and optional stopping) does not affect inference from the Bayes factor (e.g., Berger & Wolpert, 1988; Edwards et al., 1963; Lindley, 1957; Rouder, 2014; Rouder & Haaf, 2020), despite recently raised concerns (de Heide & Grünwald, 2020; Sanborn & Hills, 2014; Sanborn et al., 2014; Yu et al., 2014).

Waldian t tests require the analyst to define stopping criteria and calculate the Bayes factor repeatedly until one of the two thresholds is reached. Importantly, however, they fully preserve the assumed prior structure of Bayesian t tests. Thus, if we accept that the interpretation of the Bayes factor is unaffected by the stopping rule, Waldian t tests have the same fully Bayesian justification as Bayesian t tests. As long as sampling is terminated only upon crossing the predefined thresholds, the procedure automatically satisfies the desired frequentist properties. This does not affect, nor is it affected by, any Bayesian interpretation of the evidence provided by the particular observed sample. The Bayes factor is still a measure of statistical evidence for the specified models and it may be used to calculate posterior probabilities as well as Bayesian error probabilities. Hence, Waldian t tests provide the best of two worlds by controlling frequentist error probabilities of Bayesian t tests while preserving all of their Bayesian properties.

Simulations

We discussed the properties of Waldian t tests above based on previously presented mathematical derivations (Berger et al., 1999; Hendriksen et al., 2020; Wald, 1947). However, these derivations do not account for nuisance factors such as overshooting at the point of termination. Moreover, while the procedure controls expected error probabilities across the specified prior distributions, we know little about the error probabilities at specific values of the standardized effect size. To address these open questions, we examined the Waldian t test in a series of Monte Carlo simulations. These simulations were performed in R (R Core Team, 2021), based on publicly available functions for informed Bayesian t tests provided by Gronau et al. (2020) and closed-form expressions for Bayes factor calculation for NAP t tests provided in Pramanik and Johnson (2021). We simulated Waldian t tests for the prior specifications shown in Figure 1. Note that these priors are just convenient illustrative examples. The particular situation at hand and substantive considerations as to the meaning of the prior may require different priors. The simulation scripts as well as all simulated data are available at the accompanying Open Science Framework repository (Schnuerch et al., 2021).

Expected Error Probabilities

To examine how overshooting affects the expected error rates, we simulated data from the statistical models specified in the Waldian t tests. Under the null hypothesis, we drew random data from two normal distributions with means $\mu_x = \mu_y = 0$ and common standard deviation $\sigma = 1$. Under the alternative hypothesis, an effect size δ was randomly drawn from the specified prior in a first step. The data were then sampled from two normal distributions with means $\mu_x = \delta$ and $\mu_y = 0$. For each replication, the Bayes factor was calculated for an initial sample size of $n_x = n_y = 2$. This was subsequently increased by $+1$ in each group[2] until the Bayes factor reached one of the specified boundary values $A = (1 - \beta)/\alpha$ or $B = \beta/(1 - \alpha)$. The nominal error probabilities were systematically varied along $\alpha \in \{.005, .050\}$ and $\beta \in \{.050, .100\}$. As soon as one of the two thresholds was crossed, sampling was terminated and the respective hypothesis was accepted. If no threshold had been reached at $n_x = n_y = 25{,}000$, sampling was terminated and the hypothesis better supported by the Bayes factor was accepted. This occurred in 0.03% of the cases. For each of the 24 parameter combinations (i.e., 2 α levels $\times$ 2 β levels $\times$ 2 scenarios $\mathcal{H}_0$ and $\mathcal{H}_1 \times 3$ prior distributions under $\mathcal{H}_1$), 10,000 replications were simulated.

The results of this simulation are shown in Figure 2. Empirical error rates (i.e., proportion of replications with decisions in favor of the false hypothesis) as well as 95% CIs are displayed as a function of the true data-generating mechanism (Panel A: null hypothesis; Panel B: alternative hypothesis), nominal error rates (α and β), and specified prior settings (informed versus default versus NAP). Under the null hypothesis (Panel A), Waldian t tests reliably control Type I error rates. The proportion of erroneous decisions closely approximate the nominal α. As expected, due to overshooting, the tests (the default test in particular) are slightly conservative. This deviation is only small, however, and does not affect the statistical power of the tests: The empirical error rates under the alternative hypothesis (Panel B) almost perfectly match the nominal βs for all simulated scenarios. Thus, despite the impact of overshooting, Waldian t tests provide reliable and accurate control of expected frequentist error probabilities in practice. Table A1 contains the average sample sizes (as well as the $50th$, $75th$, and $95th$ quantile) of the simulated Waldian t tests.

Error Probabilities for Fixed Effect Sizes

In a second set of simulations, we assessed the characteristics of Waldian t tests when simulating data with a fixed effect size. To that end, we drew random data from two normal

---

[2]We chose pairwise sampling for reasons of computational efficiency. This is not necessary in practice, however. Additional observations can be drawn randomly from either group without compromising error control. Note, however, that the test may be less efficient if group sample sizes differ systematically (Schnuerch & Erdfelder, 2020).
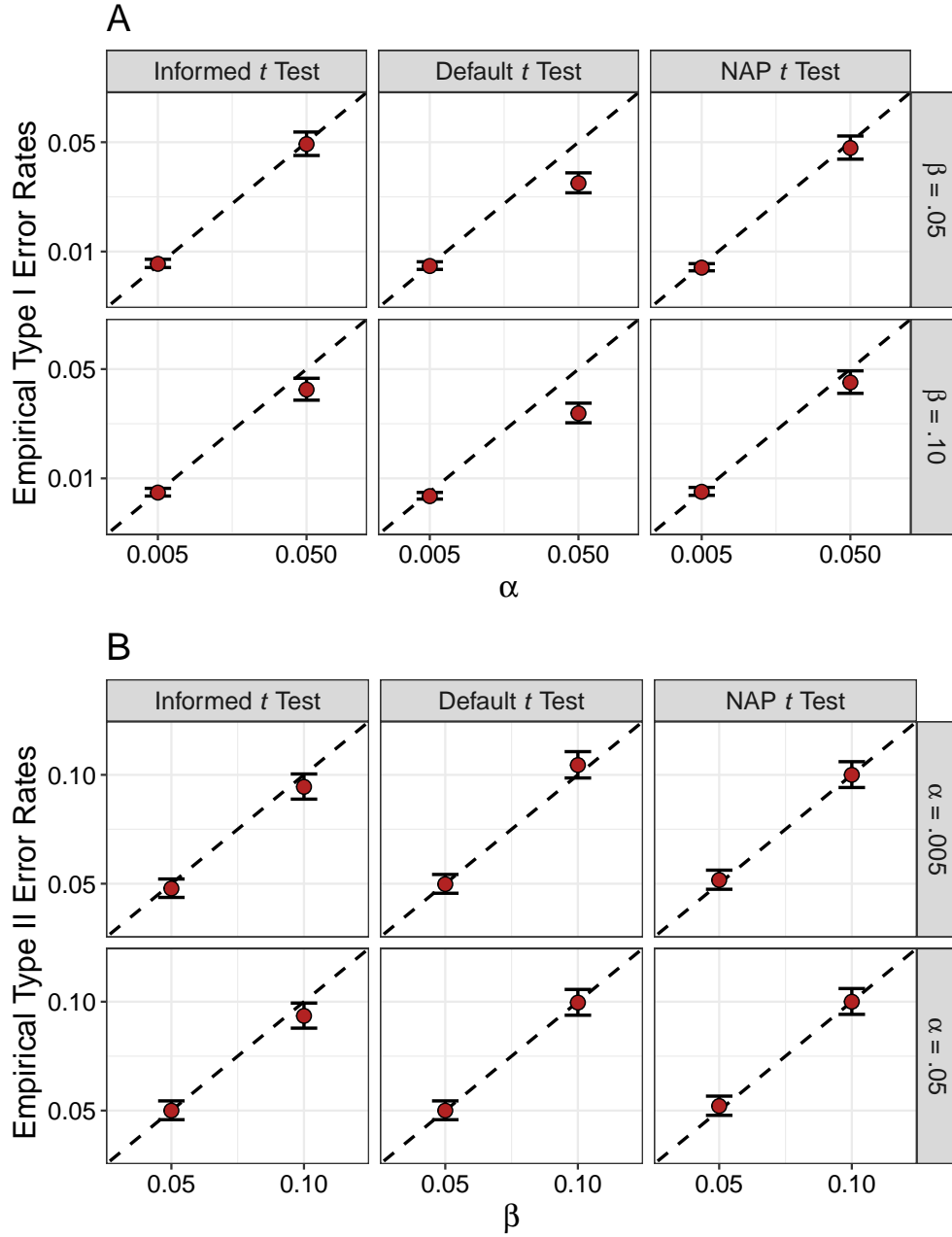
Figure 2. Empirical error rates of the Waldian t test for 10,000 replications per data point. Error bars denote 95% confidence intervals. Informed t test: t prior under the alternative ($\mu_\delta = 0.350$, $\gamma = .102$, $\kappa = 3$); Default t test: Scaled Cauchy prior under the alternative ($\mu_\delta = 0$, $\gamma = 1/\sqrt{2}$, $\kappa = 1$); NAP t test: normal moment prior under the alternative ($\mu_\delta = 0$, $\tau^2 = 0.045$, $\sigma^2 = 1$). **A**: Data generated under the null hypothesis ($\delta = 0$); **B**: Data generated under the alternative hypothesis.

distributions with common standard deviation $\sigma = 1$, $\mu_x = \delta$, and $\mu_y = 0$. We varied the effect size δ from $\delta = 0.00$ to $\delta = 0.50$ in steps of 0.05. As in the first simulation, Bayes factors were calculated for an initial sample size of $n_x = n_y = 2$. These samples were increased by $+1$ in each group until the Bayes factor reached one of the specified threshold values or the maximum sample size of $n_x = n_y = 25,000$ (which happened in 0.01% of the cases). The nominal error probabilities underlying the calculation of threshold values were held constant at $\alpha = \beta = .05$. Again, 10,000 replications per parameter combination were simulated.

For comparison, we also simulated Hajnal's t tests (Schnuerch & Erdfelder, 2020) under the same population scenarios. Whereas the null hypothesis in Hajnal's t test is identical to that in Waldian t tests (i.e., $\delta = 0.00$), a point hypothesis is specified under the alternative (i.e., $\delta = \delta_1$). Hence, it is based on the simple ratio of a non-central to a central t density. We simulated two-sided tests with the alternative hypothesis specifying very small to medium effect sizes, $\delta_1 \in \{0.10, 0.20, 0.50\}$. All simulation parameters were identical to the Waldian t test simulations, including the true effect size δ which varied from $\delta = 0.00$ to $\delta = 0.50$ and the nominal error probabilities α and β. All simulated trajectories of Hajnal's t tests reached the decision thresholds A or B before reaching the maximum sample size.

The simulation results are displayed in Figure 3. As expected, while Waldian t tests show the desired error probability $\alpha = .05$ when $\delta = 0.00$, the Type II error probability β increases monotonically with the effect size (Panel A1). For extremely small effect sizes close to zero, all tests have low statistical power. With increasing effect size, however, the power increases rapidly. For $d < 0.30$, the power of the default t test clearly exceeds that of the informed and the NAP t test. This is not surprising because the Cauchy prior under the alternative hypothesis in the default test places considerably more weight on these small effect sizes than the priors in the other two tests. The higher power comes at a cost, however: Expected sample sizes in the default test are much larger for small effect sizes than in the other tests (Panel A2). This reflects the low evidence accumulation rate induced by the high similarity of statistical models specified under $\mathcal{H}_0$ and $\mathcal{H}_1$ in the default test (Johnson & Rossell, 2010).

Hajnal's t test reliably controls error probabilities when the true effect size matches one of the hypothesized effect sizes (i.e., $\delta = 0.00$ or $\delta = \delta_1$) or when it is larger. If the true effect size is smaller than hypothesized, however, statistical power is smaller than 95%. (Panel B1). The operating characteristic (i.e., the power function) of Hajnal's t test for $\delta_1 = 0.10$ is very similar to that of the Waldian t test based on a default Cauchy prior. At the same time, expected sample sizes are much smaller (Panel B2) for small effect sizes (i.e., $\delta \leq 0.20$). In this range, Hajnal's t test based on $\delta_1 = 0.10$ is more sensitive to deviations

from the null, and evidence accumulates faster. If the true effect size is larger than expected, however, the statistical models under $\mathcal{H}_0$ and $\mathcal{H}_1$ are more difficult to distinguish, and Waldian t tests based on a default Cauchy prior become more efficient than Hajnal's t test (see also Stefan et al., 2020).

Interestingly, the operating characteristics of Waldian t tests based on the informed t prior and on the NAP are quite similar to each other and to Hajnal's t test based on $\delta_1 = 0.20$ across the simulated range of effect sizes. This similarity is not surprising when considering the range on which the respective priors place most of their weight (see Figure 1). Note, however, that the NAP places only half of its weight on positive effects (as does Hajnal's two-sided t test). The similarity between the NAP and the informed t test for the prior settings considered here can thus be expected to be even higher if the NAP was truncated at 0, that is, for a one-sided test.

Importantly, the presented simulation results depend on the specific prior settings considered in this article. For different prior distributions or threshold values based on different (expected) error probabilities, the exact results will differ. However, the simulations demonstrate that the Type II error probability $\beta(\delta)$ of a Waldian t test may vary heavily depending on the true effect size $\delta \neq 0$ that underlies a specific study. In other words, while Waldian t tests reliably control the expected error probability β given the hypothesized prior distribution of effect sizes (i.e., a weighted average of $\beta(\delta)$), they do not allow clear-cut statements about the error probability $\beta(\delta)$ for any fixed effect size δ. To control error probabilities at fixed effect sizes with maximum efficiency, Hajnal's t test is better suited (Schnuerch & Erdfelder, 2020).

Frequentist Implications of Conventional Criteria

Jeffreys (1961; see also Lee & Wagenmakers, 2013) suggested a taxonomy for the interpretation and communication of statistical evidence based on half-units of the Bayes factor on a $\log_{10}$ scale. According to this taxonomy, a Bayes factor greater than $10^{1/2} \approx 3$ denotes *moderate* evidence, greater than $10^1 = 10$ represents *strong* evidence, and greater than $10^{3/2} \approx 30$ can be interpreted as *very strong* evidence. Kass and Raftery (1995) proposed a more stringent classification based on units of twice the natural logarithm of the Bayes factor. According to their definition, the evidence indicated by the Bayes factor can be classified as *positive* (moderate) when it is between 3 and 20, *strong* when it exceeds 20, and *very strong* when it is above 150.} The specific threshold values of conventional categories like these might serve as anchors for researchers who aim at collecting sufficient evidence with a Bayesian t test to make a statistical decision. For example, a researcher could decide a priori to increase the sample size until the Bayes factor indicated *strong* evidence
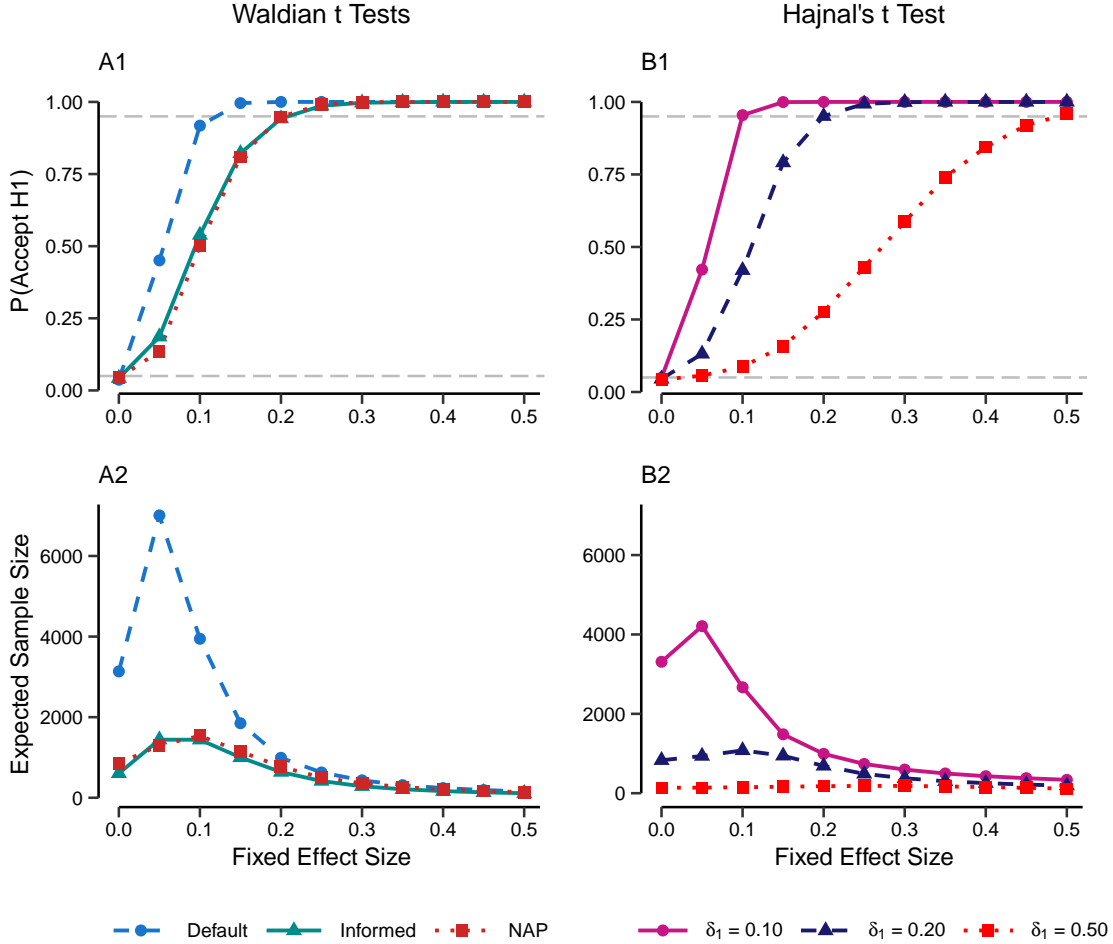
Figure 3. Properties of Waldian t tests and Hajnal's t tests based on nominal error probabilities $\alpha = \beta = .05$ for fixed effect sizes. The top row (Panels **A1** and **B1**) shows the probability to accept the alternative hypothesis; the upper and lower dashed horizontal lines denote 95% and 5% probability, respectively. The bottom row (Panels **A2** and **B2**) shows expected sample sizes. The left column (Panels **A1** and **A2**) shows results for Waldian t tests based on the following specifications: Default = scaled Cauchy prior under the alternative ($\mu_\delta = 0$, $\gamma = 1/\sqrt{2}$, $\kappa = 1$); Informed = t prior under the alternative ($\mu_\delta = 0.350$, $\gamma = .102$, $\kappa = 3$). NAP = normal moment prior under the alternative ($\mu_\delta = 0$, $\tau^2 = 0.045$, $\sigma^2 = 1$). The right column (Panels **B1** and **B2**) shows results for Hajnal's t tests based on the alternative hypothesis $\delta = \delta_1$ with $\delta_1 \in \{0.10, 0.20, 0.50\}$.

according to Jeffreys's definition in favor of a hypothesis. Accordingly, this researcher would set up a sequential Bayesian t test with upper and lower threshold values $A = 10$ and $B = 1/10$, respectively. The verbal label "strong evidence" associated with these thresholds, however, might convey misleading intuitions about the frequentist properties of the sequential procedure.

To assess the long-run error rates of a sequential Bayesian t test based on heuristic threshold values, we can use the formulas derived by Wald (1947). By solving Equation 12 for α and β, we obtain the approximate error probabilities of a sequential test with given threshold values. For a Bayesian t test with upper and lower thresholds A and B on BF_{10}, this renders

$$\alpha = \frac{1 - B}{A - B} \tag{14}$$

and

$$\beta = \frac{AB - B}{A - B}. \tag{15}$$

If the chosen threshold values are symmetric, that is, $B = 1/A$, the resulting test procedure has symmetric error probabilities as well. In this case, the above formulas reduce to

$$\alpha = \beta = \frac{1}{A + 1}. \tag{16}$$

According to Equation 16, a sequential Bayesian t test with symmetric thresholds of 10 and 1/10 is associated with frequentist error probabilities $\alpha = \beta = .09$. These Type I and Type II error rates might seem unexpectedly high when considering that the corresponding verbal label implies "strong evidence" (according to Jeffreys's taxonomy) and that statistical decisions based on a Bayes factor of 10 have typically been compared to NHST decisions based on $\alpha = .05$ (e.g., Brysbaert, 2019). Less surprisingly, employing a threshold denoting "moderate evidence" according to Jeffreys ($A = 1/B = 3$) implies even larger error rates: $\alpha = \beta = .25$. This endorses previous recommendations in the literature to avoid making decisions based on such low threshold values (Schönbrodt et al., 2017, p. 332).

Indeed, the verbal labels suggested by Kass and Raftery (1995) seem more in line with the frequentist error probabilities of a sequential procedure based on these thresholds: A sequential Bayesian t test with symmetric thresholds of 20 and 1/20, which denotes strong evidence in this definition, implies the frequentist error probabilities $\alpha = \beta = .048$. Very strong evidence, in turn, denoted by threshold values of 150 and 1/150, is then associated with $\alpha = \beta = .007$. Table 1 summarizes the nominal frequentist error rates of sequential Bayesian t tests for selected threshold values, as well as the verbal labels associated with these thresholds according to Jeffreys (1961) and Kass and Raftery (1995).

It is important to note that Equations 14 and 15 define upper bounds to error rates of the sequential procedure associated with certain thresholds A and B. These should not be confused with Bayesian error rates conditional on a particular observed result $\mathrm{BF}_{10}(N) > A$ or $\mathrm{BF}_{10}(N) < B$. Nevertheless, Equation 16 illustrates an interesting relationship between frequentist and Bayesian error rates of sequential Bayesian t tests: If the researcher terminated the sampling process as soon as $\mathrm{BF}_{10}(N)$ was exactly equal to one of the thresholds, that is, $\mathrm{BF}_{10}(N) = A$ or $\mathrm{BF}_{10}(N) = B$, the Bayesian error probabilities given 1-to-1 prior odds (see Equation 7) were identical to the frequentist error probabilities for symmetric thresholds (see Equation 16; Berger et al., 1999). Since the observed Bayes factor will typically exceed the threshold at termination (i.e., overshooting), the frequentist error probabilities of the procedure thus represent upper boundaries of the Bayesian error rates. Consequently, from a Bayesian perspective, Equations 14 and 15 are useful tools to evaluate the properties of a sequential Bayesian t test with given threshold values. Moreover, this relationship between frequentist and Bayesian error probabilities also holds for a Waldian t test with A and B chosen to satisfy certain error probabilities from a frequentist perspective. This further highlights the usefulness of Waldian t tests as a tool that combines frequentist and Bayesian desiderata.

Waldian t Tests in Practice

To facilitate the application of Waldian t tests for substantive researchers, we created an easy-to-use R Shiny web application. The app provides an intuitive graphical user interface and does not require any prior knowledge of R. It can be accessed via https://martinschnuerch.shinyapps.io/Waldian-t-Tests/, and the underlying source code is available at https://github.com/mschnuerch/waldian-t-tests.

Before data collection, the app may be used to compute the threshold values for the Waldian t test. This computation only requires the user to provide the desired error probabilities α and β. Alternatively, the threshold values A and B can be entered to calculate the associated error probabilities.

During data collection, users can repeatedly calculate and monitor the Bayes factor with the app. Before analysis, the prior on δ has to be specified. The app currently supports the specification of a scaled t distribution (which includes a scaled Cauchy as a special case), the normal moment prior proposed by Pramanik and Johnson (2021), and a normal distribution. To calculate the Bayes factor at a given stage (i.e., $\mathrm{BF}_{10}(N)$), the user must provide the observed t value as well as the group sample sizes n_x and n_y. Instead of the t value, the app also accepts group means and standard deviations. When all information has been entered, the app returns the calculated Bayes factor as well as the decision to either

Table 1

Association of threshold values and nominal error probabilities for sequential Bayesian t tests

$\mathcal{H}_1$ Threshold			$\mathcal{H}_0$ Threshold				
A	J	K&R	B	J	K&R	α	β
10	strong	moderate	10^{-1}	strong	moderate	0.091	0.091
20	strong	strong	10^{-1}	strong	moderate	0.045	0.095
30	very strong	strong	10^{-1}	strong	moderate	0.030	0.097
150	very strong	very strong	10^{-1}	strong	moderate	0.006	0.099
10	strong	moderate	20^{-1}	strong	strong	0.095	0.045
20	strong	strong	20^{-1}	strong	strong	0.048	0.048
30	very strong	strong	20^{-1}	strong	strong	0.032	0.048
150	very strong	very strong	20^{-1}	strong	strong	0.006	0.050
10	strong	moderate	30^{-1}	very strong	strong	0.097	0.030
20	strong	strong	30^{-1}	very strong	strong	0.048	0.032
30	very strong	strong	30^{-1}	very strong	strong	0.032	0.032
150	very strong	very strong	30^{-1}	very strong	strong	0.006	0.033
10	strong	moderate	150^{-1}	very strong	very strong	0.099	0.006
20	strong	strong	150^{-1}	very strong	very strong	0.050	0.006
30	very strong	strong	150^{-1}	very strong	very strong	0.033	0.006
150	very strong	very strong	150^{-1}	very strong	very strong	0.007	0.007

Note. A, B = upper and lower threshold of the sequential procedure, respectively. J = interpretation of thresholds according to Jeffreys (1961). K&R = interpretation of thresholds according to Kass and Raftery (1995). α, β = nominal Type I and Type II error probabilities associated with threshold values A and B; note that these values denote properties of the sequential procedure and do not consider effects of overshooting.

terminate and accept one of the hypotheses or to continue sampling.

To summarize, Waldian t tests are easily applied in practice. With our Shiny app, researchers can define threshold values (or error probabilities associated with fixed thresholds) and calculate Bayes factors for one-sample, two-sample, and paired Waldian t tests. The app requires only the observed t values (or group means and standard deviations). This input is easily obtained with SPSS or any freely available software such as R or JASP. When using Waldian t tests in practice, it is important to keep in mind that they—like any other sound test procedure—require specification of the models corresponding to $\mathcal{H}_0$ and $\mathcal{H}_1$ as well the parameters of the sequential procedure (i.e., α and β) *before* data collection has started. Changing the models and/or parameters during the sampling process may introduce severe biases with unknown consequences and should thus be avoided by any means. One effective

way to avoid biases is by committing to the rule that the test procedure (i.e., the prior distributions and error probabilities/threshold values) is preregistered before data collection has started (Wagenmakers et al., 2012).

Discussion

Bayesian hypothesis testing methods have become considerably popular among psychologists, fostered by advances in computational methods and persistent critiques of classical, frequentist NHST approaches to statistical inference (Heck et al., in press; Tendeiro & Kiers, 2019). One influential milestone in this process was the development of Bayesian solutions for one of the most frequent hypothesis testing scenarios, that is, Bayesian t tests (Jeffreys, 1948, 1961; see also Gönen et al., 2005; Gronau et al., 2020; Rouder et al., 2009, for recent efforts to extend and promote these tests). Bayesian t tests possess a number of attractive features. Most importantly, they allow analysts to quantify evidence in favor of the null hypothesis (Gallistel, 2009) and can be applied sequentially during the sampling process.

Notwithstanding these favorable features, standard Bayesian t tests do not provide a natural basis for controlling error probabilities. Many research contexts, however, compel researchers to choose a specific course of action based upon whether one accepts or rejects the null hypothesis (Lakens, 2021). Think of a clinical psychologist, for example, who has to decide whether or not to implement a new therapy. This decision depends on whether or not the hypothesis is accepted that the new therapy is better than the old one. Similarly, experimental psychologists might conduct a pilot study to test a specific hypothesis and decide to continue this line of research depending on accepting or rejecting the hypothesis.

A decision to accept or reject a hypothesis can always be wrong, and for a single instance there is no way to tell with absolute certainty that no mistake has been made. From the perspective of cumulative science, however, it is crucial that the overall proportion of erroneous decisions is sufficiently low (Lakatos, 1978; Neyman, 1977). Good frequentist properties raise confidence in the result of statistical tests (Sanborn et al., 2014). Specifically, they indicate how severely (i.e., how critically) a hypothesis has been tested (Mayo & Spanos, 2006; Westermann & Hager, 1986). Therefore, we believe that these properties should be considered and controlled also in the context of Bayesian t tests, at least in situations that compel the analyst to take a certain course of action. The issue has been addressed before and different procedures have been proposed to control error probabilities in the context of Bayesian t tests (e.g., Berger et al., 1999; Berger et al., 1994; Gu et al., 2016; Schönbrodt & Wagenmakers, 2018; Stefan et al., 2020). With this article, we complement these suggestions: The Waldian t test is a straightforward unification of Bayesian t tests with sequential probability ratio tests. It allows for simple and efficient control of frequentist

error probabilities while preserving the prior structure of the Bayesian test. Moreover, in this context, the frequentist error probabilities serve as upper bounds for the Bayesian error probabilities at the point of termination. Thus, Waldian t tests satisfy both frequentist and Bayesian desiderata by combining their respective advantages.

Limitations

Statistical procedures are tools suited for different situations with different advantages and limitations. There is no magic tool that fits all purposes. Waldian t tests are no exception. We want to address some of their limitations here.

Waldian t tests inherit the SPRT's procedural assumption that sampling can be continued until a decision threshold is reached. This assumption is shared among non-truncated sequential tests and it is important in the derivation of their long-run properties. Although such sequential procedures are much more efficient on average than procedures based on fixed samples, there is no guarantee that the test will terminate at or before reaching a certain sample size (Schnuerch & Erdfelder, 2020; Schönbrodt et al., 2017). Concluding from simulations, the risk is small that the required sample size becomes unfeasibly large. Nevertheless, this feature of Waldian t tests may limit their applicability to scenarios in which the specification of a definite upper bound to the sample size beforehand is not pivotal.

If a definite upper bound to the sample size is mandatory, a standard Neyman-Pearson test based on a fixed sample size derived from an a priori power analysis might be more appropriate. Alternatively, one could use a sequential design with some maximum sample size, for example, group sequential designs (Lakens, 2014) or the independent segments procedure (Miller & Ulrich, 2020; see also Erdfelder & Schnuerch, 2021). To allow for proper error control, however, these approaches require that both the null and the alternative hypothesis be specified as simple point hypotheses. If the analyst's expectation is best represented by a prior distribution, a Bayesian t test is more appropriate. In this case, the necessary fixed sample size and critical value such that the procedure satisfies certain error rates may be determined by a simulation-based design analysis (Stefan et al., 2019). Alternatively, the recently proposed modified sequential probability ratio test (Pramanik et al., 2021) allows for a sequential test of a fixed null hypothesis against a default alternative hypothesis based on uniformly most powerful Bayesian tests (Johnson, 2013) with a pre-defined maximum sample size. This modified sequential test represents an efficient hypothesis testing procedure when the alternative cannot be specified explicitly and when an upper bound to the sample size is required.

Another important point to keep in mind is that the specified Type II error probability β in a Waldian t test is an expected probability. As we illustrated in our simulations (Figure

3), the error probability for a specific parameter value, $\beta(\delta)$, will vary considerably across the parameter space Ω_1 and thus, deviate from this expectation in general. As we argue above, Waldian t tests will nevertheless provide reasonable frequentist error probability control if the prior represents meaningful properties of δ (e.g., a random effect).

A different approach is to construct a test of the smallest effect size of interest. In this case, one would specify a theoretically or practically meaningful effect size $\delta_{\min} \in \Omega_1$ for which the requirement is imposed that the test has error probability less or equal to β for any δ greater than or equal to $\delta_{\min}$. A Bayesian (or Waldian) t test would no longer be appropriate for this scenario. However, a different sequential procedure such as Hajnal's t test (Schnuerch & Erdfelder, 2020) or a fixed-sample Neyman-Pearson t test based on the point alternative hypothesis $\mathcal{H}_1\colon \delta = \delta_{\min}$ will satisfy the error probability requirement, that is, $P(\text{reject } \mathcal{H}_1;\ \delta) \leq \beta,\ \forall \delta \in \{\delta\colon\ \delta \geq \delta_{\min}\}$.

Conclusion

Different research questions require different statistical answers. While some questions may be appropriately addressed by continuous measures of statistical evidence, others require binary decisions with controlled error probabilities. With Waldian t tests, we promote a procedure that combines the advantages of Bayesian t tests, which aim at quantifying evidence, with that of sequential probability ratio tests, which provide an efficient means to control error probabilities. What is more, the basic idea underlying Waldian t test can of course be generalized to Bayes factors outside the t test framework, such as Bayes factors aiming at model selection or assessing order constraints (for a review of applications, see Heck et al., in press). This, however, exceeds the scope of this article and needs to be worked out in more detail in the future.

It is not our intention to suggest that psychologists should abandon hypothesis testing based on Bayesian tests, Neyman-Pearson tests, NHST, or other extant procedures. Instead, we endorse the mindful selection of statistical tools suited for the research question at hand and the given practical constraints (e.g., the costs of data collection). We believe that for many basic and applied psychological research questions, Waldian t tests will constitute a valuable addition to the set of available tools.

References

Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. *Nature*, *567*(7748), 305–307. https://doi.org/10.1038/d41586-019-00857-9

Armitage, P., McPherson, C. K., & Rowe, B. C. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society. Series A (General)*, *132*(2), 235–244. https://doi.org/10.2307/2343787

Aust, F., & Barth, M. (2020). Papaja: Create APA manuscripts with R Markdown (R package version 0.1.0.9942).

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin*, *66*(6), 423–437. https://doi.org/10.1037/h0020412

Barnard, G. A. (1949). Statistical inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, *11*(2), 115–149. https://doi.org/10.1007/978-1-4613-8505-9

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

Berger, J. O. (2003). Could Fisher, Jeffreys and Neyman have agreed on testing? *Statistical Science*, *18*(1), 1–32. https://doi.org/10.1214/ss/1056397485

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

Berger, J. O., & Bayarri, M. J. (2004). The interplay of Bayesian and frequentist analysis. *Statistical Science*, *19*(1), 58–80. https://doi.org/10.1214/088342304000000116

Berger, J. O., Boukai, B., & Wang, Y. (1997). Unified frequentist and Bayesian testing of a precise hypothesis. *Statistical Science*, *12*, 133–160. https://doi.org/10.1214/ss/1030037904

Berger, J. O., Boukai, B., & Wang, Y. (1999). Simultaneous Bayesian-frequentist sequential testing of nested hypotheses. *Biometrika*, *86*(1), 79–92. https://doi.org/10.1093/biomet/86.1.79

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

Berger, J. O., & Wolpert, R. L. (1988). *The likelihood principle: A review, generalizations, and statistical implications*. Hayward, CA, Institute of Mathematical Statistics.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK, John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470743386

Bredenkamp, J. (1972). *Der Signifikanztest in der psychologischen Forschung [The test of significance in psychological research].* Darmstadt, Germany, Akademische Verlagsgesellschaft.

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? a tutorial of power analysis with reference tables. *Journal of Cognition*, *2*(16), 1–38. https://doi.org/10.5334/joc.72

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology*, *65*(3), 145–153. https://doi.org/10.1037/h0045186

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (Second). Hillsdale, NJ, Erlbaum.

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist*, *49*(12), 997–1003. https://doi.org/10.1037/0003-066X.49.12.997

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

de Heide, R., & Grünwald, P. D. (2020). Why optional stopping can be a problem for Bayesians. *Psychonomic Bulletin & Review.* https://doi.org/10.3758/s13423-020-01803-x

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

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers*, *28*(1), 1–11. https://doi.org/10.3758/BF03203630

Erdfelder, E., & Schnuerch, M. (2021). On the efficiency of the independent segments procedure: A direct comparison with sequential probability ratio tests. *Psychological Methods*, *26*(4), 501–506. https://doi.org/10.1037/met0000404

Gallistel, C. R. (2009). The importance of proving the null. *Psychological Review*, *116*(2), 439–453. https://doi.org/10.1037/a0015251

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., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, *66*(1), 8–38. https://doi.org/10.1111/j.2044-8317.2011.02037.x

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, NY, Erlbaum.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*(5), 587–606. https://doi.org/10.1016/j.socec.2004.09.033

Gillett, R. (1994). Post hoc power analysis. *Journal of Applied Psychology*, *79*(5), 783–785. https://doi.org/10.1037/0021-9010.79.5.783

Gönen, M., Johnson, W. O., Lu, Y., & Westfall, P. H. (2005). The Bayesian Two-Sample *t* Test. *The American Statistician*, *59*(3), 252–257. https://doi.org/10.1198/000313005X55233

Gronau, Q. F., Ly, A., & Wagenmakers, E.-J. (2020). Informed Bayesian *t* -Tests. *The American Statistician*, *74*(2), 137–143. https://doi.org/10.1080/00031305.2018.1562983

Gu, X., Hoijtink, H., & Mulder, J. (2016). Error probabilities in default Bayesian hypothesis testing. *Journal of Mathematical Psychology*, *72*, 130–143. https://doi.org/10.1016/j.jmp.2015.09.001

Hajnal, J. (1961). A two-sample sequential t-test. *Biometrika*, *48*(1/2), 65–75. https://doi.org/10.2307/2333131

Heck, D. W., Boehm, U., Böing-Messing, F., Bürkner, P. C., Derks, K., Dienes, Z., Fu, Q., Gu, X., Karimova, D., Kiers, H., Klugkist, I., Kuiper, R. M., Lee, M. D., Leenders, R., Leplaa, H. J., Linde, M., Ly, A., Meijerink-Bosman, M., Moerbeek, M., . . . Hoijtink, H. (in press). A review of applications of the bayes factor in psychological research. *Psychological Methods*. https://doi.org/10.31234/osf.io/cu43g

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando, Academic Press.

Hendriksen, A., de Heide, R., & Grünwald, P. (2020). Optional stopping with Bayes factors: A categorization and extension of folklore results, with an application to invariant situations. *Bayesian Analysis*, *16*(3), 961–989. https://doi.org/10.1214/20-BA1234

Hoijtink, H., van Kooten, P., & Hulsker, K. (2016). Why Bayesian psychologists should change the way they use the bayes factor. *Multivariate Behavioral Research*, *51*(1), 2–10. https://doi.org/10.1080/00273171.2014.969364

JASP Team. (2020). JASP (Version 0.14.1)[Computer software].

Jeffreys, H. (1948). *Theory of probability* (2nd Ed.). Oxford, Oxford University Press.

Jeffreys, H. (1961). *Theory of probability* (3rd Ed.). Oxford, Oxford University Press.

Jeon, M., & De Boeck, P. (2017). Decision qualities of Bayes factor and p value-based hypothesis testing. *Psychological Methods*, *22*(2), 340–360. https://doi.org/10.1037/met0000140

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Johnson, V. E. (2013). Uniformly most powerful Bayesian tests. *The Annals of Statistics*, *41*(4), 1716–1741. https://doi.org/10.1214/13-AOS1123

Johnson, V. E., & Rossell, D. (2010). On the use of non-local prior densities in Bayesian hypothesis tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *72*(2), 143–170. https://doi.org/10.1111/j.1467-9868.2009.00730.x

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Lakatos, I. (1978). *The methodology of scientific research programmes.* Cambridge ; New York, Cambridge University Press.

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, *44*(7), 701–710. https://doi.org/10.1002/ejsp.2023

Lakens, D. (2021). The practical alternative to the *p* value is the correctly used *p* value. *Perspectives on Psychological Science*, *16*(3), 639–648. https://doi.org/10.1177/1745691620958012

Lang, A.-G. (2017). Is intermediately inspecting statistical data necessarily a bad research practice? *The Quantitative Methods for Psychology*, *13*(2), 127–140. https://doi.org/10.20982/tqmp.13.2.p127

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

Lindley, D. V. (1957). A statistical paradox. *Biometrika*, *44*(1-2), 187–192. https://doi.org/10.1093/biomet/44.1-2.187

Ly, A., Verhagen, J., & Wagenmakers, E.-J. (2016). Harold Jeffreys's default Bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology*, *72*, 19–32. https://doi.org/10.1016/j.jmp.2015.06.004

Mayo, D. G. (1996). *Error and the growth of experimental knowledge.* Chicago, University of Chicago Press.

Mayo, D. G. (2018). *Statistical inference as severe testing: How to get beyond the statistics wars.* Cambridge ; New York, NY, Cambridge University Press.

Mayo, D. G., & Spanos, A. (2006). Severe testing as a basic concept in a Neyman–Pearson philosophy of induction. *The British Journal for the Philosophy of Science*, *57*(2), 323–357. https://doi.org/10.1093/bjps/axl003

Miller, J., & Ulrich, R. (2020). A simple, general, and efficient method for sequential hypothesis testing: The independent segments procedure. *Psychological Methods.* https://doi.org/10.1037/met0000350

Morey, R. D., Romeijn, J.-W., & Rouder, J. N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*, *72*, 6–18. https://doi.org/10.1016/j.jmp.2015.11.001

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

Neyman, J. (1977). Frequentist probability and frequentist statistics. *Synthese*, *36*(1), 97–131. https://doi.org/10.1007/BF00485695

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *231*(694-706), 289–337. https://doi.org/10.1098/rsta.1933.0009

Pramanik, S., & Johnson, V. E. (2021). *Efficient alternatives for Bayesian hypothesis tests in psychology.*

Pramanik, S., Johnson, V. E., & Bhattacharya, A. (2021). A modified sequential probability ratio test. *Journal of Mathematical Psychology*, *101*, 102505. https://doi.org/10.1016/j.jmp.2021.102505

R Core Team. (2021). R: A language and environment for statistical computing. *Vienna, Austria.*

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

Rouder, J. N., & Haaf, J. M. (2020). *Optional stopping and the interpretation of the Bayes factor* (Preprint). PsyArXiv. https://doi.org/10.31234/osf.io/m6dhw

Rouder, J. N., & Morey, R. D. (2017). Teaching Bayes' theorem: Strength of evidence as predictive accuracy. *The American Statistician*, *73*, 186–190. https://doi.org/10.1080/00031305.2017.1341334

Rouder, J. N., Morey, R. D., & Wagenmakers, E.-J. (2016). The interplay between subjectivity, statistical practice, and psychological science. *Collabra*, *2*, 1–12. https://doi.org/10.1525/collabra.28

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. https://doi.org/10.3758/PBR.16.2.225

Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin*, *57*(5), 416–428. https://doi.org/10.1037/h0042040

Rushton, S. (1950). On a sequential t-test. *Biometrika*, *37*(3/4), 326–333. https://doi.org/10.2307/2332385

Rushton, S. (1952). On a two-sided sequential t-test. *Biometrika*, *39*(3/4), 302. https://doi.org/10.2307/2334026

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

Sanborn, A. N., Hills, T. T., Dougherty, M. R., Thomas, R. P., Yu, E. C., & Sprenger, A. M. (2014). Reply to Rouder (2014): Good frequentist properties raise confidence. *Psychonomic Bulletin & Review*, *21*(2), 309–311. https://doi.org/10.3758/s13423-014-0607-4

Schnuerch, M., & Erdfelder, E. (2020). Controlling decision errors with minimal costs: The sequential probability ratio t test. *Psychological Methods*, *25*(2), 206–226. https://doi.org/10.1037/met0000234

Schnuerch, M., Erdfelder, E., & Heck, D. W. (2020). Sequential hypothesis tests for multinomial processing tree models. *Journal of Mathematical Psychology*, *95*, 102326. https://doi.org/10.1016/j.jmp.2020.102326

Schnuerch, M., Heck, D. W., & Erdfelder, E. (2021). Supplemental material for "Waldian t tests: Sequential Bayesian t tests with controlled error probabilities". Retrieved from https://osf.io/z5vsy/.

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, *25*, 128–142. https://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*(2), 322–339. https://doi.org/10.1037/met0000061

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

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*. https://doi.org/10.3758/s13428-018-01189-8

Stefan, A. M., Schönbrodt, F. D., Evans, N. J., & Wagenmakers, E.-J. (2020). *Efficiency in sequential testing: Comparing the sequential probability ratio test and the sequential Bayes factor test* (Preprint). PsyArXiv. https://doi.org/10.31234/osf.io/ry4fw

Tendeiro, J. N., & Kiers, H. A. L. (2019). A review of issues about null hypothesis Bayesian testing. *Psychological Methods*, *24*(6), 774–795. https://doi.org/10.1037/met0000221

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. https://doi.org/10.1027/2151-2604/a000319

van Ravenzwaaij, D., & Wagenmakers, E.-J. (in press). Advantages masquerading as 'issues' in Bayesian hypothesis testing: A commentary on Tendeiro and Kiers (2019). *Psychological Methods.*

Vanpaemel, W. (2010). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology, 54*(6), 491–498. https://doi.org/10.1016/j.jmp.2010.07.003

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review, 14*, 779–804. https://doi.org/10.3758/BF03194105

Wagenmakers, E.-J., & Gronau, Q. F. (2018). Error rate schmerror rate. *Bayesian Spectactles.*

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7*(6), 632–638. https://doi.org/10.1177/1745691612463078

Wald, A. (1947). *Sequential analysis.* New York, Wiley.

Westermann, R., & Hager, W. (1986). Error probabilities in educational and psychological research. *Journal of Educational Statistics, 11*(2), 117–146. https://doi.org/10.3102/10769986011002117

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. https://doi.org/10.1080/14786442108633773

Yu, E. C., Sprenger, A. M., Thomas, R. P., & Dougherty, M. R. (2014). When decision heuristics and science collide. *Psychonomic Bulletin & Review, 21*(2), 268–282. https://doi.org/10.3758/s13423-013-0495-z

Appendix

Table A1

Average Sample Size (and Quantiles) of Simulated Waldian t Tests

α	β	True state	Informed t test	Default t test	NAP t test
0.005	0.05	$\mathcal{H}_0$	699 (430,708,1847)	3675 (2434,3496,8949)	954 (694,974,2135)
		$\mathcal{H}_1$	568 (316,566,1619)	1200 (102,574,5047)	723 (416,812,2325)
	0.10	$\mathcal{H}_0$	326 (198,352,926)	911 (602,856,2151)	544 (402,560,1192)
		$\mathcal{H}_1$	427 (271,490,1187)	565 (108,518,2307)	593 (380,680,1739)
0.050	0.05	$\mathcal{H}_0$	608 (382,614,1568)	3178 (2194,3121,7649)	852 (652,891,1845)
		$\mathcal{H}_1$	313 (166,331,998)	810 (64,380,3765)	445 (238,548,1463)
	0.10	$\mathcal{H}_0$	269 (180,310,733)	798 (542,764,1887)	491 (376,512,1080)
		$\mathcal{H}_1$	236 (150,272,687)	329 (60,368,1405)	352 (238,422,1016)

Note. α = nominal Type I error probability; β = nominal Type II error probability; True state = statistical model underlying data generation; Informed t test: t prior under the alternative ($\mu_\delta = 0.350$, $\gamma = .102$, $\kappa = 3$); Default t test: Scaled Cauchy prior under the alternative ($\mu_\delta = 0$, $\gamma = 1/\sqrt{2}$, $\kappa = 1$); NAP t test: Non-local alternative prior ($\mu_\delta = 0$, $\sigma^2 = 1$, $\tau^2 = 0.045$). Values in parentheses represent the $50th$, $75th$, and $95th$ quantile, respectively.