Improving Statistical Practice in Psychological Research: Sequential Tests of Composite Hypotheses

MARTIN SCHNÜRCH

Inaugural Dissertation
Submitted in partial fulfillment of the requirements for the degree of Doctor of Social Sciences in the DFG Research Training Group "Statistical Modeling in Psychology" at the University of Mannheim

Thesis Defense:
February 26, 2020

Supervisor:
Prof. Dr. Edgar Erdfelder

Dean of the School of Social Sciences:
Prof. Dr. Michael Diehl

Thesis Reviewers:
Prof. Dr. Thorsten Meiser
Prof. Dr. Rolf Ulrich

Defense Committee:
Prof. Dr. Edgar Erdfelder
Prof. Dr. Thorsten Meiser
Prof. Dr. Rolf Ulrich

For my family

Contents

Summary

Statistical hypothesis testing is an integral part of the scientific process. When employed to make decisions about hypotheses, it is important that statistical tests control the probabilities of decision errors. Conventional procedures that allow for error-probability control have limitations, however: They often require extremely large sample sizes, are bound to tests of point hypotheses, and typically require explicit assumptions about unknown nuisance parameters. As a consequence, the issue of proper error-probability control has frequently been neglected in statistical practice, resulting in a widespread reliance on questionable statistical rituals.

In this thesis, I promote an alternative statistical procedure: the sequential probability ratio test (SPRT). In three articles, I implement, further develop, and examine three extensions of the SPRT to common hypothesis-testing situations in psychological research. In the first project, I show that the SPRT substantially reduces required sample sizes while reliably controlling error probabilities in the context of the common t-test situation. In a subsequent project, I seize on the SPRT to develop a simple procedure that allows for statistical decisions with controlled error probabilities in the context of Bayesian t tests. Thus, it allows for tests of distributional hypotheses and combines the advantages of frequentist and Bayesian hypothesis tests. Finally, I apply a procedure for sequential hypothesis tests without explicit assumptions about unknown nuisance parameters to a popular class of stochastic measurement models, namely, multinomial processing tree models. With that, I demonstrate how sequential analysis can improve the applicability of these models in substantive research.

The procedures promoted herein do not only extend the SPRT to common hypothesis-testing situations, they also remedy a number of limitations of conventional hypothesis tests. With my dissertation, I aim to make these procedures available to psychologists, thus bridging the gap between the fields of statistical methods and substantive research. Thereby, I hope to contribute to the improvement of statistical practice in psychology and help restore public trust in the reliability of psychological research.

Articles

This dissertation is the result of research conducted in the context of the research training group "Statistical Modeling in Psychology" (SMiP). It is based on three articles, two of which have been published and one has been submitted for publication. In line with the core idea of the research training group, each article can be located within a cuboid defined by the three dimensions of SMiP's research agenda: statistical techniques, model families, and application fields.

In the main text, I discuss how each of the articles relates to these dimensions and how they build upon each other. Thereby, I outline the unifying framework of my dissertation project, which aims to improve statistical practice in psychological research by bridging the gap between methodological and substantive research.

By further developing, implementing, and demonstrating three approaches to test composite hypotheses with sequential techniques, I show how these methods can improve the efficiency of hypothesis tests (Article I), combine the advantages of Bayesian and frequentists tests (Article II), and facilitate the application of more complex stochastic models (Article III). In that, I hope that my dissertation makes a significant contribution to improving the methods employed to approach substantive psychological research questions.

Article I

Schnuerch, M., & Erdfelder, E. (2019). Controlling decision errors with minimal costs: The sequential probability ratio *t* test. *Psychological Methods*. Advance online publication. http://doi.org/10.1037/met0000234

Article II

Schnuerch, M., Heck, D. W., & Erdfelder, E. (2019). *Waldian t tests for accepting and rejecting the null hypothesis with controlled error probabilities.* Manuscript submitted for publication.

Article III

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

And so these men of Indostan
    Disputed loud and long,
Each with his own opinion
    Exceeding stiff and strong,
Though each was partly in the right,
    And all were in the wrong.

from *The Blind Men and the Elephant*
by J. G. Saxe

1 Introduction

Every year, societies dedicate a considerable amount of their resources to scientific research. In 2018, the German Federal Ministry of Education and Research's budget amounted to around 18.1 billion Euros (Bundesministerium für Bildung und Forschung, 2019). The German Research Society, funded mainly by the federal ministry, granted nearly 200 million Euros to research projects in social and behavioral sciences alone (Deutsche Forschungsgemeinschaft, 2019). Considering that this is but a small portion of what is spent on research on a global scale, these figures impressively illustrate the societal impact of scientific research.

Not least because of this, researchers have a responsibility not to squander the resources with which they have been trusted. Recent findings that a great number of seemingly well-established results failed to replicate, however, have cast serious doubt on the extent to which psychological researchers fulfill this responsibility. Replicability is a hallmark of scientific progress (e.g., Hempel & Oppenheim, 1948; Platt, 1964), and reports on low replicability rates in psychology have marked the dawn of a far-reaching confidence crisis (Earp & Trafimow, 2015; Ioannidis, 2005; Maxwell, Lau, & Howard, 2015; Open Science Collaboration, 2015; Pashler & Wagenmakers, 2012).

To restore public trust, many have argued for profound changes in research practices (e.g., Asendorpf et al., 2013; Begley & Ioannidis, 2015; Benjamin et al., 2018; Chalmers et al., 2014; Chalmers & Glasziou, 2009; Cumming, 2014; Dienes, 2016; Ioannidis et al., 2014; Munafò et al., 2017; Nosek, Spies, & Motyl, 2012). A particular focus in this discussion has been on the statistical methods employed to test hypotheses. The dominant procedure in psychology (and many other fields) is typically referred to as *null-hypothesis significance testing* (NHST). As a somewhat logically incoherent compound of the seminal theories of significance testing by Fisher (1935a) and statistical decision making by Neyman and Pearson (1933), NHST has been criticized for decades (Bakan, 1966; Bredenkamp, 1972; Cohen, 1994; Gigerenzer, 1993, 1998, 2004; Goodman, 1993; Rozeboom, 1960; Sedlmeier, 1996; Wagenmakers, 2007). Broader awareness of its shortcomings has only recently emerged, however, as the replication crisis fostered strong pleas for its renunciation (e.g., Cumming, 2014; Dienes, 2011).

Among other things, attention has been drawn to the need for strict control of probabilities of statistical decision errors. Figure 1 illustrates the influence of the Type-I error probability α, that is, the probability to reject the null hypothesis when it is *true*, and the statistical power $1 - \beta$, that is, the probability to reject the null hypothesis when it is *false*, on the replicability rate. The replicability rate denotes the proportion of successful replications of statistically significant results, that is, results which led to a rejection of the null hypothesis. Obviously, this rate is not only a function of error probabilities but also of the proportion of false null hypotheses. This proportion is commonly referred to as the *base rate* of true alternative hypotheses. For a given base rate, replicability typically increases, ceteris paribus, with increasing statistical power and decreasing Type-I error probability. Figure 1 depicts the replicability rate as a function of α and $1 - \beta$ for a base rate of .10, which is a reasonable estimate for a range of research areas in experimental psychology (Miller & Ulrich, 2016, 2019).

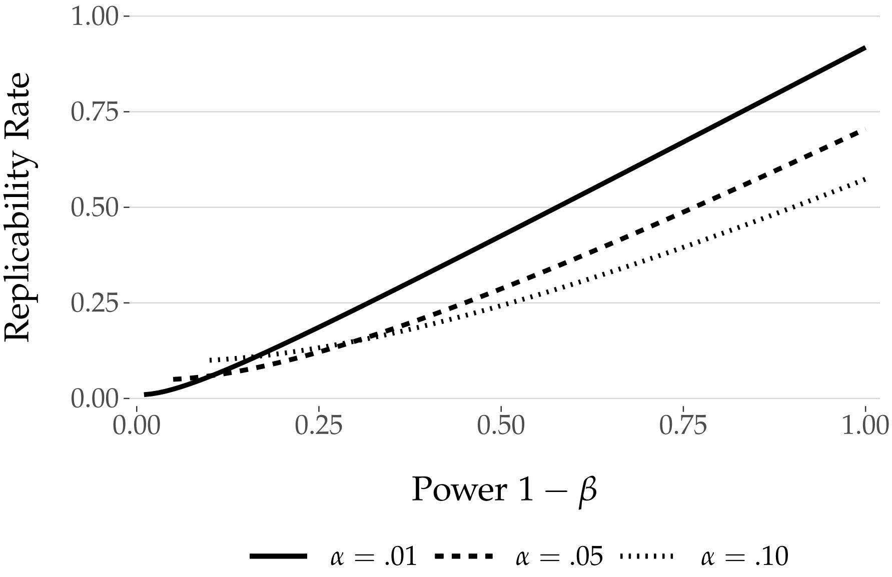

Figure 1: Replicability rate as a function of Type-I error rate α and statistical power $1 - \beta$. Base rate of alternative hypothesis = .10.

According to the statistical guidelines of the Psychonomic Society, "it is important to address the issue of statistical power. ... Studies with low statistical power produce inherently ambiguous results because they often fail to replicate" (Psychonomic Society, 2019). Despite such pleas, however, statistical power has often been neglected. This is partly due to the sample-size requirements of common procedures that allow for error-probability control (i.e., Neyman-Pearson procedures). Especially when to-be-detected effect sizes are small, Neyman-Pearson tests require extremely large sample sizes (Erdfelder, Faul, & Buchner, 1996). As a consequence, most experiments feature notably smaller sample sizes, resulting in an average power of around $1 - \beta = .50$ in prototypical journal publications (Cohen, 1962; Sedlmeier & Gigerenzer, 1989).

In my thesis, I promote an alternative statistical method that has been developed more than 70 years ago: the sequential probability ratio test (SPRT; Wald, 1947). In Neyman-Pearson tests, the sample size required to satisfy certain error probabilities is defined by an a priori power analysis (Cohen, 1988; Faul, Erdfelder, Buchner, & Lang, 2009). Sequential tests, in contrast, dispense with the requirement of a fixed sample size. Instead, the data are continuously monitored during the sampling process until a predefined criterion is met. Thus, a defining feature of sequential tests is that sampling is terminated as soon as the data show a compelling result. Therefore, sequential analyses require on average substantially smaller samples than conventional statistical techniques. At the same time, they allow for error-probability control (Wetherill, 1975).

Apart from applications in clinical research (Proschan, Lan, & Wittes, 2006), sequential analyses have largely been ignored in psychology over the past decades (Botella, Ximénez, Revuelta, & Suero, 2006; Lang, 2017). This might be surprising, given the many beneficial properties of sequential tests. Considering the above-mentioned responsibility not to waste valuable resources, one might even identify an "ethical obligation" of researchers to make use of the most efficient methods available (Lakens, 2014, p. 701).

One of the reasons for the widespread neglect of sequential methods has been their mathematical sophistication (Botella et al., 2006). Until the late 20th century, high-performance computers were not available. Thus, in order to apply sequential analyses researchers had to rely on complex mathematical calculations and approximations. Nowadays, in contrast, statistical analyses are easily conducted with standard statistical software and, thus, mathematical complexity is no longer a limitation.

Another reason that has limited the applicability of sequential tests is that they are typically designed for so-called *simple* hypotheses. A hypothesis is simple when the values of all parameters of the statistical model that defines the probability distribution of the data are uniquely determined. If more than a single value for each parameter is consistent with the hypothesis, it is not simple but *composite* (Wald, 1947). Let $\theta_1, ..., \theta_K$ denote the population parameters of the probability distribution of some random variable X, that is, $X \sim f(x|\theta_1, ..., \theta_K)$, with $x, \theta_k \in \mathbb{R}$. If $K = 2$, for example, the hypothesis that $\theta_1 = 0$ and $\theta_2 = 0$ is simple. In contrast, the hypotheses that $\theta_1 = 0$ and $\theta_2 \neq 0$, or $\theta_1 = \theta_2$ are composite, because they are consistent with infinitely many values of θ_2. To summarize, a hypothesis is simple when it is consistent with exactly one point in the K-dimensional parameter space. Otherwise, it is composite.

Consider the common case of testing a hypothesis on the mean of some normally distributed random variable. A hypothesis $\mu = \mu_0$ would only be simple if the scale of the random variable (i.e., the population variance σ^2) was known. This is typically not the case, thus, σ^2 is a so-called *nuisance parameter* and many test procedures do not apply. The problem of composite hypotheses is not only an issue for sequential tests, but for all procedures that are based on simple hypotheses—such as, for example, the Neyman-Pearson procedure.

Most research questions imply composite hypotheses. As a consequence, procedures that are restricted to simple hypotheses are of limited practical use for substantive researchers. Thus, it is not surprising that the SPRT has not been widely used in psychology. To improve its applicability, such that substantive researchers can benefit from its many advantages, it is important to consider the SPRT in more realistic scenarios.

In this thesis, I study three approaches to extend SPRTs to the case of composite hypotheses. In the first article (Schnuerch & Erdfelder, 2019), I implement and examine the properties of a sequential *t* test in a simulation study. The test is based on a method by D. R. Cox (1952b) that requires a jointly sufficient set of estimators for the unknown parameters, such that a test can be constructed on transformations of the observations, whose distribution no longer depends on the unknown nuisance parameters.

In the second article (Schnuerch, Heck, & Erdfelder, 2019), I develop a sequential design for Bayesian *t* tests such that the procedure allows for statistical decisions with controlled error probabilities. The procedure is a simple extension of the SPRT, based on a suggestion by Wald (1947) tailored to alternative hypotheses that do not put restrictions on the exact value of the parameter of interest.

In the last article (Schnuerch, Erdfelder, & Heck, 2020), I seize on a method suggested by D. R. Cox (1963), where a sequential test in the presence of nuisance parameters is constructed based on asymptotic maximum-likelihood theory. I use this method to develop sequential hypothesis tests on parameters of a popular class of stochastic models for discrete data, namely, multinomial processing tree (MPT) models.

In the following, I discuss each article in more detail and elaborate on the general framework in which they are connected, that is, sequential probability ratio tests of composite hypotheses. Moreover, I put this framework into the context of the cuboid model representing the three dimensions of SMiP's research agenda: In Part I, I address the statistical-techniques dimension and how the SPRT increases the efficiency and reliability of statistical hypothesis testing (Articles I and II). In Part II, I bridge the gap to the second dimension, that is, the model-families dimension, by demonstrating how sequential tests can facilitate the applicability of MPT models (Article III). In Part III, I address the application-fields dimension by outlining with concrete examples how the methods I promote can benefit substantive research. Finally, I discuss limitations and future directions.

The field of psychology has yet to see the full ramifications of the replication crisis. With my dissertation, I hope to contribute to overcoming this crisis by making sequential hypothesis tests available to substantive researchers and improving statistical practice. However, whereas I am convinced that questionable statistical rituals have played their part in the development of the crisis, they certainly are not the sole cause of it. The research process comprises many elements and statistical hypothesis testing is only one of them. To fully understand the mechanisms underlying the replication crisis, all elements have to be considered in interaction, not in isolation. This is beyond the scope of this dissertation, but necessary for the field of psychology as a whole. Otherwise, our conclusions will be as misguided as those of the blind men from Indostan in J. G. Saxe's famous poem: In order to find out what an elephant is like, each of the men feels some part of it. One is feeling its side, one its leg, another its trunk, and so on. Then, each convinced that he has discovered the truth, they argue whether the elephant is like a wall, a tree, or a snake—when, undoubtedly, integrating their experiences would have brought them closer to the truth.

2 Part I: Statistical Techniques

> There are no routine statistical questions, only questionable statistical routines.
>
> *(D. R. Cox)*

2.1 Current Statistical Practice

Hypothesis testing is commonly viewed as a decision-making problem: Given some hypothesis $\mathcal{H}$ and observed data X, do we accept the hypothesis or do we reject it? From the perspective of an empirical science, critically testing the empirical predictions derived from a theory is an integral part of the research process. An integral part of hypothesis testing, in turn, is the use of statistical methods.

According to Ronald Fisher (1922), "the object of statistical methods is the reduction of data" (p. 311) to a smaller, tangible quantity that captures all the essential information. This is done by constructing a hypothetical distribution of samples drawn from some population, which can be described by a small number of parameters and from which we assume our data to be a random sample. The statistical process can thus be reduced to three essential problems: (1) *Specification*, that is, the choice of the statistical model representing the hypothetical population; (2) *Estimation*, that is, calculation of statistics that are used as estimates for the unknown population parameters; and (3) *Distribution*, that is, the consideration of the distribution of the calculated statistics across all possible samples.

Based on this notion, Fisher (1935a) developed his theory of significance testing. The first step of Fisher's procedure includes the specification of the statistical model underlying the data and the formulation of a simple null hypothesis on the parameter(s) of interest, $\mathcal{H}_0$: $\theta = \theta_0$. Let $x^n = (x_1, ..., x_n)$ denote a sample of n observations and $t(x^n)$ the relevant test statistic computed from the data. The test of significance is then based on the computation of the p value, denoting the conditional probability of the observed or a more extreme test statistic under the null hypothesis. For a one-tailed test, this means

$$p = P(T(x^n) \geq t(x^n) | \theta_0). \tag{2.1}$$

Fisher suggested that this value be interpreted as the *significance level*, with small values denoting statistical evidence against the null hypothesis (Barnard, 1947; Berger, 2003). This reasoning has been termed "Fisher's disjunction" (Rouder, Morey, Verhagen, Province, & Wagenmakers, 2016): A small p value indicates that either a rare event has occurred or the null hypothesis is false. Depending on the situation at hand and the observed level of significance, this reasoning may warrant either rejecting the null hypothesis or maintaining it (Barnard, 1947).

Despite its influence, Fisher's theory sparked trenchant critique. Many statisticians argued that the disjunction, as a probabilistic version of the *modus tollens*, was logically invalid (Cohen, 1994; Royall, 1997). Among its fiercest critics was Jerzy Neyman, who also challenged Fisher's concept of *fiducial probability*, that is, the argument that objective probability distributions for the unknown population parameters θ can be deduced from the data alone (Fisher, 1935b; Kalbfleisch & Sprott, 1967). In fact, the argument is hard to reconcile with the classical, frequentist view of probabilities (Neyman, 1941). Although Fisher shared this classical view (see Fisher, 1922) and repeatedly argued that the fiducial argument had been misinterpreted and misrepresented (e.g., Fisher, 1935b, 1955), Neyman remained that Fisher's theory lacked proper, frequentist justification, and set out to develop his own theory of hypothesis testing (Berger, 2003).

In collaboration with Egon Pearson, Neyman developed a most influential theory, in which the test of a hypothesis is constructed as a decision rule to decide between two possible courses of action. The aim of this rule is to govern behavior such that, in the long run, the proportion of erroneous decisions can be controlled (Neyman & Pearson, 1933). The justification for this procedure is termed the *frequentist principle*: If a statistical procedure is applied repeatedly, the overall rates of decision errors should be equal to the average reported error probabilities (Berger, 2003; Neyman, 1977).

A crucial element in the Neyman-Pearson theory is the notion that the test of a hypothesis always requires the specification of an (exact) alternative hypothesis. This can be illustrated by a simple example (see Pollard & Richardson, 1987): Assume we are meeting a random person and we want to test the following hypothesis,

$$\mathcal{H}\text{: The person is a U.S. citizen.} \tag{2.2}$$

In the course of testing this hypothesis, we might observe that the person is president of the United States. Under our hypothesis, that is, given that the person we have met is a citizen of the United States, this person being the president is an extremely unlikely event (as of November 24, 2019, the probability is approximately $1/330,000,000$; U.S. Census Bureau, 2019). Following Fisher's logic, we might thus conclude that the hypothesis is false and should be rejected. Obviously, our conclusion is misguided, as we failed to consider the probability of the observed event *given that the hypothesis is false*. Observing an event that is unlikely under one hypothesis does not imply that it is more likely under any alternative hypothesis. If the person is not a U.S. citizen, being president is not only unlikely but simply impossible. Hence, our observation, however unlikely under the hypothesis, is unequivocal evidence that it is, in fact, true. This example illustrates the importance of considering alternative hypotheses.

Let the null and the alternative hypothesis be denoted by $\mathcal{H}_0$: $\theta = \theta_0$ and $\mathcal{H}_1$: $\theta = \theta_1$, respectively. A statistical test may then warrant one of two mutually exclusive courses of action: Accept $\mathcal{H}_0$ or $\mathcal{H}_1$ and reject the respective other one. Based on this decision, two possible errors can occur, namely, accepting $\mathcal{H}_1$ when $\mathcal{H}_0$ is true and vice versa. The first one is referred to as Type-I error and the second as Type-II error, with α and β commonly denoting the respective probabilities of the statistical procedure to commit either one.

The Neyman-Pearson procedure to control these error probabilities is based on the *likelihood ratio*. This ratio quantifies how much more likely the data have occurred under one hypothesis relative to the other. Let $f(x^n|\theta_i)$ denote the probability density function of an observed sample of size n, conditional on the parameter (or parameter vector) θ_i under hypothesis i, $i = 0,\ 1$.[1] The probability density is proportional to the likelihood, $f(x^n|\theta_i) \propto \mathcal{L}(\theta_i; x^n)$, thus, the likelihood ratio is typically expressed as the ratio of probability densities:

$$LR_n = \frac{f(x^n|\theta_1)}{f(x^n|\theta_0)}. \tag{2.3}$$

This ratio measures how well one hypothesis predicted the data relative to the other hypothesis. Relative predictive accuracy, in turn, represents the statistical evidence that the data provide for each hypothesis relative to the other (Royall, 1997). In order to construct a test procedure that controls the error probabilities α and β, Neyman and Pearson suggested to use the following decision rule: Reject $\mathcal{H}_0$ if

$$LR_n \geq c, \tag{2.4}$$

where c is defined a priori, and accept it otherwise. Let Ω denote the set of all possible samples of size n, that is, $x^n \in \Omega$, $\forall x^n$, and Ω^c the critical region, that is, the set of all samples of size n for which Inequality 2.4 is satisfied. To achieve the desired Type-I error probability, c is chosen such that, if the the null hypothesis is true, the relative size of the critical region is equal to α. Moreover, it is chosen such that Ω^c consists of those elements of Ω for which the likelihood ratio LR_n takes on its maximum values. Thus, the probability that $x^n \in \Omega^c$ if the alternative hypothesis is true, that is, the power $1 - \beta$, will be maximized. In other words, there is no alternative critical region for which the

[1] The random variable X can be either continuous or discrete, in which cases $f(.)$ denotes the probability density or the probability mass, respectively. Without loss of generality, I will treat X as a continuous variable in what follows.

power would be larger. Thus, the procedure is a most powerful test of $\mathcal{H}_0$ versus $\mathcal{H}_1$ for a given α and fixed sample size. This property is described in the *Neyman-Pearson lemma* (Neyman & Pearson, 1933; Royall, 1997).

This result marks an important milestone in the theory of hypothesis testing (Wald, 1947). For any given (simple) hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$, the power of the test is a function of α and the sample size. Thus, the error probabilities of the procedure can be controlled explicitly by choosing a certain sample size. This is known as an a priori *power analysis* (Cohen, 1988, 1992; Faul et al., 2009).

Neyman-Pearson tests and power analyses have served as nominal standards for statistical tests in psychology and many other scientific fields. Common practice, however, often deviates from these norms (Erdfelder et al., 1996): For decades, the de-facto standard in psychological research has been null-hypothesis significance testing, a hybrid procedure composed of elements of both Fisher's as well as Neyman and Pearson's theories (Bredenkamp, 1972; Goodman, 1993; Wagenmakers, 2007). Like Fisher's theory, it only considers the null hypothesis but no specific alternative. At the same time, like Neyman-Pearson tests, it is used as a decision-making procedure with controlled Type-I error probability α, sometimes in combination with post-hoc assessments of effect sizes and statistical power (Gigerenzer, 2004).

As a compound of two rivaling approaches, NHST is a theoretically inconsistent hybrid procedure. Despite its pitfalls, however, it has become a pervasive statistical ritual. If there is anything that Neyman and Fisher could have agreed on, despite their fundamental differences and personal disputes, it would have been the illegitimacy of that statistical hybrid (Berger, 2003; Gigerenzer, 1993, 1998, 2004).

In recent years, awareness of the shortcomings of NHST has spread and fostered calls for a paradigm shift toward other methods (e.g., Cumming, 2014; Dienes, 2011; Rouder, Morey, & Wagenmakers, 2016; Wagenmakers, 2007). Among other things, the replication crisis has underlined the need for reliable error-probability control of statistical tests (e.g., Psychonomic Society, 2019). Neyman-Pearson tests, however, as the common procedure that allows for error-probability control, often require extremely large sample sizes. This may have partly encouraged researchers to neglect power analyses and to rely on NHST instead. Thus, not surprisingly, the average power of statistical tests reported in the field of psychology has typically been unacceptably low (Cohen, 1962; Erdfelder et al., 1996; Sedlmeier & Gigerenzer, 1989), indicating the urgent need for efficient alternatives. One promising alternative, and the main topic of this thesis, is *sequential analysis*.

2.2 The Sequential Probability Ratio Test

A typical assumption of conventional statistical tests is that the sample size is fixed, and the tests' properties are defined in reference to the set of possible outcomes of repetitions with the same sample size (Barnard, 1949). In sequential analysis, in contrast, the sample size is not defined a priori. Instead, the data are sampled sequentially, that is, at any new step of the sampling process a decision is made to either terminate (and accept one of the hypotheses) or continue sampling. This process is reiterated until the first decision is made and sampling stops. Thus, by implication, the sample size is a random variable that depends on the sequence of observations.

In the 1940's, Abraham Wald introduced one of the first formal theories of sequential analysis, the *sequential probability ratio test* (SPRT; Wald, 1945, 1947). The SPRT is based on the same test statistic as the Neyman-Pearson procedure, namely, the likelihood ratio (see Equation 2.3). As before, let $f(x^n|\theta)$ denote the probability density function of an observed sample of size n and θ the true parameter (or parameter vector) defining this distribution. For the test of the two simple hypotheses $\mathcal{H}_0$: $\theta = \theta_0$ and $\mathcal{H}_1$: $\theta = \theta_1$, the likelihood ratio (LR_n) is computed for any integral value of n, starting at $n = 1$. At each stage of this procedure, one of the following three decisions is made:

$$\begin{aligned} &\text{1) Accept } \mathcal{H}_1 \text{ and reject } \mathcal{H}_0 \text{ when } LR_n \geq A;\\ &\text{2) Accept } \mathcal{H}_0 \text{ and reject } \mathcal{H}_1 \text{ when } LR_n \leq B;\\ &\text{3) Sample a new independent observation } x_{n+1} \text{ when } B < LR_n < A. \end{aligned} \tag{2.5}$$

It is straightforward to prove that the statistical procedure defined by Equations 2.3 and 2.5 will terminate with probability 1 (see Wald, 1947, Appendix A.1). Thus, according to the nomenclature suggested by Kendall and Stuart (1969), the SPRT is a *closed* sequential scheme, as opposed to *open* procedures that may, potentially, continue indefinitely without a decision.[2] Based on this, the determination of the decision boundaries A and B such that the test satisfies certain error probabilities α and β is straightforward.

Let x_i^n denote a sample that leads to the acceptance of $\mathcal{H}_i$, $i = 0,\ 1$, that is,

$$B < \frac{f(x_i^{n-1}|\theta_1)}{f(x_i^{n-1}|\theta_0)} < A, \tag{2.6}$$

and

$$\frac{f(x_1^n|\theta_1)}{f(x_1^n|\theta_0)} \geq A \tag{2.7}$$

[2] The terminology in the statistical literature appears to be somewhat inconsistent at this point. Some authors use the term *closed* to describe test procedures with an exact upper limit on the sample size. Following Kendall and Stuart (1969, p. 593), however, I will refer to these tests as *truncated*.

or

$$\frac{f(x_0^n|\theta_1)}{f(x_0^n|\theta_0)} \leq B, \tag{2.8}$$

respectively. Thus, by definition, any sample x_1^n satisfies the following inequality,

$$f(x_1^n|\theta_1) \geq A \cdot f(x_1^n|\theta_0), \tag{2.9}$$

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 a sample x_1^n is at least A times larger under $\mathcal{H}_1$ than under $\mathcal{H}_0$. The probability to obtain a sample x_1^n, in turn, is equivalent to the probability to accept $\mathcal{H}_1$. Because the SPRT eventually terminates with either accepting $\mathcal{H}_1$ or $\mathcal{H}_0$, this implies that the probability to accept $\mathcal{H}_1$ is at least A times larger under $\mathcal{H}_1$ than under $\mathcal{H}_0$. In the usual notation, the former is defined as the statistical power $1-\beta$ and the latter as the Type-I error probability α, hence, $1-\beta \geq A\alpha$. Following the same logic for x_0^n, we see that $\beta \leq B(1-\alpha)$, and thus,

$$A \leq \frac{1-\beta}{\alpha} \tag{2.10}$$

and

$$B \geq \frac{\beta}{1-\alpha}. \tag{2.11}$$

Inequalities 2.10 and 2.11 indicate that upper and lower limits for A and B are given by $(1-\beta)/\alpha$ and $\beta/(1-\alpha)$, respectively. Wald (1947) showed that treating these inequalities as equalities to define A and B "cannot result in any appreciable increase in the value of either α or β" (p. 46). In fact, as the likelihood ratio will typically have exceeded the boundary at the point of termination (a phenomenon called *overshooting*), the resulting error rates of the sequential procedure will be lower than the nominal α and β. Wald conjectured that the resulting decrease in efficiency is negligible and, thus, suggested that for practical purposes the SPRT based on Equations 2.3 and 2.5 be performed with $A = (1-\beta)/\alpha$ and $B = \beta/(1-\alpha)$. The resulting SPRT is a test with approximate strength (α, β).

The SPRT marked an important milestone in the theoretical development of sequential analysis (Wetherill, 1975). Of particular interest is the test's *optimum character*: Wald and Wolfowitz (1948) proved that, among all tests with the same strength (α, β), the SPRT requires on average the fewest observations. Let $E_\theta(N|S)$ denote the average sample size N for a sequential test S when θ is the parameter of interest. A test S' is called optimum if, for any alternative test S of equal strength, $E_{\theta_i}(N|S') \leq E_{\theta_i}(N|S)$, $i = 0, 1$. This property has been proven for the SPRT for the case of testing a simple null hypothesis against a simple alternative (Matthes, 1963; Wald & Wolfowitz, 1948). It is important to note, however, that the proof only holds if the true θ is equal to θ_0 or θ_1 (Wetherill, 1975).

Apart from its efficiency, the SPRT has a number of desirable properties. Unlike classical Neyman-Pearson or Fisherian tests, for example, the SPRT does not require assumptions about the distribution of the test statistic. The decision thresholds are functions of α and β only. Thus, it is not necessary to assume a distribution of the likelihood ratio under the null hypothesis in order to choose a critical value that satisfies the required error probabilities (Wald, 1947).

To compute the likelihood ratio, however, the SPRT does require the likelihood function to be completely specified under each hypothesis. This limits the general theory of the SPRT, as well as analytical solutions for functions describing the test procedure's properties, to simple hypotheses. As argued above, however, hypotheses are typically composite. Apart from the mathematical complexity of calculating the likelihood ratio without the availability of computers, this limitation may explain the striking lack of SPRT applications in the field of psychology (Botella et al., 2006; Lang, 2017; Wetherill, 1975).

In the years following its introduction, several extensions of the SPRT have been proposed to adapt the general theory to the case of composite hypotheses. As I show in my thesis, these extensions not only widen the scope of the SPRT, they also contribute to overcoming some pervasive problems that practitioners typically face in statistical hypothesis testing. With these extensions, the SPRT may become an important alternative to classical test procedures, especially in light of the reproducibility crisis, and a helpful step toward better statistical practice in psychological research.

2.3 Hajnal's *t* Test

Schnuerch, M., & Erdfelder, E. (2019). Controlling decision errors with minimal costs: The sequential probability ratio *t* test. *Psychological Methods*. Advance online publication. http://dx.doi.org/10.1037/met0000234

One of the arguably most common statistical tests in psychological research is concerned with hypotheses on means (or mean differences) of normally distributed variables. Assume a one-sample test of the hypothesis $\mathcal{H}_0$: $\mu = \mu_0$ versus $\mathcal{H}_1$: $\mu = \mu_1$. Let $f(x|\mu,\, \sigma^2)$ denote the probability density of an observed value x. Only if the population variance σ^2 was known, the hypotheses would be simple and the application of the SPRT would be straightforward. If σ^2 is an unknown nuisance parameter, however, the hypotheses are composite and the likelihood function (as well as the likelihood ratio) is not fully specified.

One way to cope with the problem of nuisance parameters is to transform the observed sequence to one that no longer depends on the unknown parameter(s). This idea was first formulated by Armitage (1947) and later by Barnard (1949), who suggested to use instead of the sample observations $x_1, ..., x_n$ the corresponding t statistics $t_2, ..., t_n$ computed from the observed data to construct the likelihood ratio,

$$LR_n = \frac{f(t_2, ..., t_n | \mathcal{H}_1)}{f(t_2, ..., t_n | \mathcal{H}_0)}. \tag{2.12}$$

The t distribution no longer depends on the unknown population variance. It is fully defined by the degrees of freedom df_n and the noncentrality parameter Δ_i corresponding to hypothesis $\mathcal{H}_i$ at the n^{th} stage.

Since the sequence of t values is not composed of independent elements, this ratio can be complex to calculate. According to a theorem presented by D. R. Cox (1952b), however, the likelihood function in Equation 2.12 factorizes into

$$f(t_2, ..., t_n | \mathcal{H}_i) = f(t_2, ..., t_n | df_n, \Delta_i) = f(t_n | df_n, \Delta_i) \cdot f(t_2, ..., t_{n-1} | t_n), \tag{2.13}$$

the last term of which no longer depends on the hypothesis. Thus, the ratio at the n^{th} stage can be reduced to the ratio of the densities of t_n under each hypothesis (Rushton, 1950, 1952),

$$LR_n = \frac{f(t_n | df_n,\ \Delta_1)}{f(t_n | df_n,\ \Delta_0)}. \tag{2.14}$$

As a straightforward application of this result, Hajnal (1961) showed that a sequential two-sample t test for hypotheses on mean differences between two independent populations can be constructed based upon the same principle. In Schnuerch and Erdfelder (2019), we implemented Hajnal's t test in the statistical computing environment R (R Core Team, 2019) and examined its properties by means of simulations. Additionally, we compared the test with the classical Neyman-Pearson test, the group sequential (GS) test (Proschan et al., 2006), and sequential Bayes factors (SBFs; Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, 2017) in terms of error-probability control, efficiency, and robustness against violations of assumptions.

Being a sequential approach, the GS design allows for interim analyses during the sampling process. Unlike Hajnal's t test, however, it is based on a fixed number of planned stops, including a number of interim and one final test. The sample sizes at each stop are defined a priori. For example, a researcher might plan to inspect the data after $n = 25$, 50, and 75 observations, and perform a final test at $N_{\max} = 100$ observations. At each step, the fixed-sample test statistic is computed and compared with critical values that are calculated for each stop based on linear spending functions for the overall error rates of the procedure, α and β (Lakens, 2014). Thus, in each stage, the test can either terminate and accept $\mathcal{H}_0$ or $\mathcal{H}_1$, or sampling is continued until the next stage is reached. In the final stage, the test will accept one of the hypotheses. Thus, the procedure is truncated, since the required sample size can never exceed N_{max}.

The GS design allows for explicit control of error probabilities for two competing hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$. At the same time, it requires on average fewer observations than classical Neyman-Pearson tests with the same error probabilities. Although it is based on normally distributed test statistics, the GS design can also be applied in the context of t tests, since the t distribution approximates the normal distribution with increasing sample size (Lang, 2017; Schnuerch & Erdfelder, 2019).

The SBFs are rooted in the Bayesian framework, where the method of choice for hypothesis testing and model comparison is the Bayes factor (Jeffreys, 1961; Wrinch & Jeffreys, 1921). The Bayes factor denotes the multiplicative factor by which subjective belief is updated in light of the data (see Section 2.4). Like the test statistic in the SPRT, it is a likelihood ratio and, thus, a measure of evidence in the data for one hypothesis vis-à-vis the other (Kass & Raftery, 1995):

$$BF_{10} = \frac{f(x^n|\mathcal{H}_1)}{f(x^n|\mathcal{H}_0)}. \tag{2.15}$$

The likelihoods specified in this ratio are so-called *marginal* likelihoods. In Bayesian statistics, uncertainty about the exact values of unknown parameters is expressed by means of prior distributions on the parameters. These distributions represent the subjective belief about how plausible different parameter values are. The probability density of observed data under hypothesis $\mathcal{H}_i$, the marginal likelihood, is then obtained by integrating across the prior distributions $\pi_{\mathcal{H}_i}(\theta)$,

$$f(x^n|\mathcal{H}_i) = \int_{\Theta_{\mathcal{H}_i}} f_{\mathcal{H}_i}(x^n|\theta)\pi_{\mathcal{H}_i}(\theta)\, d\theta. \tag{2.16}$$

In Equation 2.16, $\Theta_{\mathcal{H}_i}$ is the parameter space defined by hypothesis i and $f_{\mathcal{H}_i}(x^n|\theta)$ denotes the probability density of the data for a specific point θ in $\Theta_{\mathcal{H}_i}$. Thus, the Bayes factor denotes a weighted average likelihood ratio for all possible parameter values (Morey & Rouder, 2011). As such, it crucially depends on the specified prior distributions. We focused on the arguably most prominent Bayes factor specification introduced by Rouder, Speckman, Sun, Morey, and Iverson (2009), the Bayesian t test. Here, the hypotheses are represented by the prior distributions on the standardized mean difference δ (Cohen's d). Under the null hypothesis, the prior is a point mass on $\delta = 0$. Under the alternative, in contrast, the prior is a Cauchy distribution, a heavy-tailed distribution whose shape is defined by a scale parameter r. For $r = 1$, the Cauchy is a t distribution with one degree of freedom.

The SBFs are based on the sequential calculation of the Bayes factor until it reaches some predefined threshold. Unlike in the SPRT, however, this threshold is not based on considerations of required error probabilities. Therefore, to keep the error rates of all four test procedures constant in order to compare their efficiency, we simulated the SBF design for a wide range of population scenarios, varying the true effect size δ, the scale parameter r of the Cauchy prior, and the decision threshold. For each parameter combination, we then used the empirical error rates of the SBFs to construct a classical Neyman-Pearson t test, a GS test, and Hajnal's t test with the same error probabilities.

We replicated the results of previous simulation studies in showing that the sequential designs are on average substantially more efficient than the classical, fixed-sample procedure (Schönbrodt et al., 2017). What is more, we showed that Hajnal's t test is even more efficient than the GS and SBF tests with corresponding error rates (Schnuerch & Erdfelder, 2019, Figure 2).

Moreover, we investigated the impact of violations of basic assumptions on the test procedures' performance. If the effect sizes specified under the hypotheses are much larger than the true effect, or if the true effect is random rather than fixed, Hajnal's t test is negatively affected in terms of error rates and efficiency. At the same time, in a balanced design, Hajnal's t test is quite robust against violations of normality and homogeneity assumptions, much more so than the other sequential designs.

To summarize, with the first article, we showed that the SPRT can be extended based on D. R. Cox's (1952b) method to the case of composite hypotheses for the t test. We implemented the SPRT in standard statistical software and examined the properties of one SPRT t test, Hajnal's t test, by means of simulations. Thus, we demonstrated empirically the advantages of the SPRT compared with classical as well as other sequential designs. This underlines our argument that the SPRT constitutes an attractive alternative to current statistical practice by combining reliable error-probability control with high efficiency.

A possible limitation of the non-truncated sequential design is the risk to end up with extremely large samples. The general theory of the SPRT is based on continuous sampling until a threshold is reached, and the error rates of this procedure critically hinge on this assumption. Although our simulations showed that the risk is small (typically, around 90% of the simulations terminate with a sample size smaller than the corresponding Neyman-Pearson test), the situation might occur in single instances. If the sequential procedure is terminated before reaching a boundary, a decision must not be made. Otherwise, the error rates of such a procedure are unknown and potentially much larger than intended.

Another limitation is that Hajnal's t test is bound to the test of point hypotheses. For many situations, the specification of point hypotheses is reasonable, for example, as lower limits for a substantively meaningful effect size that the test should detect with power $\geq 1 - \beta$ (Psychonomic Society, 2019; Schulz & Grimes, 2005). Sometimes, however, there might be no reason to assume a certain, fixed effect size. If, for example, the theory at test predicts the absence of an effect, any $\delta \neq 0$ might be interpreted as contradicting the hypothesis. In this case, it might be more appropriate to specify a distribution (representing a random effect) rather than a point under the alternative hypothesis.

The Bayes factor can be used to test such hypotheses, since it is based on prior distributions. As a continuous measure of evidence, however, it does not constitute a natural basis for statistical decisions. Thus, if the Bayes factor is employed to accept or reject hypotheses, the long-run error rates of that procedure are not controlled explicitly. In the second article of my dissertation, I developed a sequential design, based on an extension of the SPRT suggested by Wald (1947), which allows for statistical decisions based on the Bayes factor with controlled error probabilities.

2.4 Waldian *t* Tests

Schnuerch, M., Heck, D. W., & Erdfelder, E. (2019). *Waldian t tests for accepting and rejecting the null hypothesis with controlled error probabilities.* Manuscript submitted for publication.

In what is now one of the most-often cited articles published in *Psychonomic Bulletin & Review*, Rouder et al. (2009) proposed *Bayesian t tests* as an alternative to NHST for accepting and rejecting the null hypothesis. As mentioned above, Bayesian *t* tests use the Bayes factor to quantify statistical evidence for competing hypotheses. The fundamental assumption at the heart of Bayesian statistics is that probability represents subjective belief: How certain are we that something is true or will happen (Etz & Vandekerckhove, 2018)? Based on this, we can put probabilities on single events (*What's the probability that Germany will win the FIFA World Cup?*), hypotheses (*What's the probability that standing next to a giant box makes people more creative, i.e., 'think outside the box'?*; Lee & Wagenmakers, 2013), or parameter values.

From this perspective, hypothesis testing means updating the belief that a hypothesis is true by looking at data. A principled way how one should update subjective belief in light of data is given by Bayes' well-known theorem (i.e., the law of inverse probability). The Bayes factor is a consequence of a straightforward application of this theorem to the case of two competing hypotheses. In order to arrive at the relative belief in two competing hypotheses *after* seeing the data (posterior odds), the relative belief *before* seeing the data (prior odds) is multiplied with the relative accuracy with which the hypotheses predicted the data:

$$\underbrace{\frac{P(\mathcal{H}_1|\text{data})}{P(\mathcal{H}_0|\text{data})}}_{\text{Posterior odds}} = \underbrace{\frac{f(\text{data}|\mathcal{H}_1)}{f(\text{data}|\mathcal{H}_0)}}_{\text{Bayes factor}} \cdot \underbrace{\frac{P(\mathcal{H}_1)}{P(\mathcal{H}_0)}}_{\text{Prior odds}} \tag{2.17}$$

This multiplicative updating factor is the Bayes factor. As explained in the previous section, it is a ratio of marginal likelihoods. As such, it measures the statistical evidence in the data for one hypothesis relative to the other (Kass & Raftery, 1995; Rouder & Morey, 2017). It has frequently been argued that measuring evidence and updating subjective belief to posterior odds or probabilities reflects the core aim of statistical inference or even science in general (e.g., Dienes, 2011; Edwards, Lindman, & Savage, 1963; Morey, Romeijn, & Rouder, 2016; Rouder, Morey, & Wagenmakers, 2016; Rozeboom, 1960), and Bayesian hypothesis tests have gained notable attraction in psychological research (Tendeiro & Kiers, 2019).

In practice, however, hypothesis testing often resembles a decision-making process (cf. Berger, 2006; Schnuerch & Erdfelder, 2019). Consider experimental psychologists conducting a pilot study. Based on their decision to accept or reject the hypothesis of interest in that study, they might decide to continue or abandon this particular line of research. In the same vein, clinical researchers might decide to implement a new treatment based on their decision to accept or reject the hypothesis that it is better than the old treatment. As in these examples, many situations compel researchers to dichotomize the continuous Bayes factor into discrete regions of acceptance or rejection (Jeon & De Boeck, 2017). For these situations, it is vital to consider and control the long-run rates of incorrect decisions of the statistical procedure (Sanborn & Hills, 2014; Sanborn et al., 2014). For the standard Bayesian t test, however, there is no means to control decision-error probabilities explicitly. As a remedy, we propose a simple extension of Bayesian t tests based on the SPRT (Schnuerch et al., 2019; see also Berger, Boukai, & Wang, 1999).

In the Bayesian t test for two independent samples, the data from the two groups are modeled as normally distributed with means $\mu \pm \delta\sigma/2$ and common variance σ^2. Thus, the statistical model comprises three parameters: the grand mean μ, the standardized effect size δ, and the variance σ^2, all of which are typically unknown. Therefore, prior distributions are defined for the unknown parameters, denoting subjective belief about the plausibility of possible values.

Following suggestions by Jeffreys (1961) and Zellner and Siow (1980), the Bayesian t test employs a so-called *JZS prior* on the unknown parameters. For the nuisance parameters μ and σ^2, the priors are non-informative and equivalent under both $\mathcal{H}_0$ and $\mathcal{H}_1$, rendering their influence on the resulting Bayes factor negligible. Thus, the statistical hypotheses actually tested are represented by the priors on the standardized effect size

δ (Rouder et al., 2009).

Under the null hypothesis, the prior is a point mass corresponding to the classical, simple null hypothesis that the population means are identical, $\mathcal{H}_0$: $\delta = 0$. Under the alternative, in contrast, the prior is a Cauchy distribution representing the hypothesis $\mathcal{H}_1$: $\delta \sim \text{Cauchy}(r)$. It is easy to see that this hypothesis is composite because the Cauchy has support over the entire real line. Thus, infinitely many $\delta \in \mathbb{R}$ are consistent with the hypothesis.

Error-probability control in the classical Neyman-Pearson sense requires a procedure for which

$$P(\text{accept } \mathcal{H}_i|\mathcal{H}_i) = \begin{cases} 1-\alpha & (i=0) \\ 1-\beta & (i=1) \end{cases}, \tag{2.18}$$

where $P(\text{accept } \mathcal{H}_i|\mathcal{H}_i)$ denotes the probability to accept $\mathcal{H}_i$ if it is, in fact, true. For $\mathcal{H}_1$ in the Bayesian t test, this means that we require the test procedure to have an average Type-II error probability β across all possible values of δ. For this situation, Wald (1947) outlined a simple solution in the SPRT framework (see also Berger et al., 1999).

If we can specify the plausibility of different values of δ by means of some weight function $\omega(\delta)$ that integrates to one across the parameter space Δ_1, then the composite hypothesis $\mathcal{H}_1$ on δ is conceptually equivalent to a simple hypothesis $\mathcal{H}_1^*$ on the probability distribution of the data, that is,

$$\mathcal{H}_1^* : \; x \sim f_1(x) = \int_{\delta \in \Delta_1} f(x|\delta)\; \omega(\delta)\; d\delta. \tag{2.19}$$

Wald (1947) showed that an SPRT for the simple hypothesis $\mathcal{H}_0$ against the simple alternative $\mathcal{H}_1^*$ with $A = (1-\beta)/\alpha$ and $B = \beta/(1-\alpha)$ will have Type-I error probability equal to α and Type-II error probability equal to

$$\int_{\delta \in \Delta_1} \beta(\delta)\; \omega(\delta)\; d\delta = \beta, \tag{2.20}$$

where $\beta(\delta)$ denotes the Type-II error probability for a specific value of δ. The Cauchy prior distribution on the effect size in the Bayesian t test is conceptually equivalent to the weight function $\omega(\delta)$. Thus, Wald's (1947) result applies: An SPRT with the Bayes factor as likelihood ratio and threshold values A and B will approximately satisfy the error requirements defined in (2.18). We implemented this procedure and demonstrated by means of simulations that it does indeed satisfy the error-probability requirements as specified. What is more, this result holds for any proper prior distribution, not just that suggested by Rouder et al. (2009).

We refer to this newly developed sequential design for the Bayesian t test as *Waldian t test*. It is an analytically derived framework that allows for statistical decisions with controlled error probabilities conditional on the hypotheses tested with the Bayes factor. Thereby, it combines the beneficial properties of classical, frequentist tests with those of Bayesian tests: While it allows for error-probability control, it preserves the fully Bayesian interpretation of the Bayes factor as a measure of evidence for the specified statistical models and updating factor of relative belief.

It is important to note that for a composite hypothesis represented by a prior distribution, Waldian t tests control the *average* Type-II error probability. This implies that the error probability of the procedure will not be constant for any parameter value $\delta \in \Delta_1$. From a classical frequentist point of view, a prior distribution (i.e., weight function) represents a random effect, that is, true variation in parameter values or effect sizes across single experiments. This assumption is reasonable, for example, if the theory at test predicts the absence of an effect ($\delta = 0$) and under the alternative hypothesis ($\delta \neq 0$), the prior distribution represents approximately the range of possible, differently weighted effect sizes in the field. In this case, controlling the average error probability is in line with the classical error requirements as specified in (2.18).

If a hypothesis test is required with an upper-limit error probability for any δ in a range of possible values, a test based on simple hypotheses is more appropriate. In this case, we would need to define a minimum relevant effect size $\delta_{\text{min}} \in \Delta_1$. Hajnal's t test or a Neyman-Pearson t test based on the point alternative hypothesis $\mathcal{H}_1$: $\delta = \delta_{\text{min}}$ would then satisfy the error requirement, that is, for both test procedures, $P(\text{reject } \mathcal{H}_1|\delta) \leq \beta$, $\forall \delta \in \{\Delta_1 | \delta \geq \delta_{\text{min}}\}$ (Schnuerch & Erdfelder, 2019). If no minimum relevant effect size can be specified, however, as in the case of a substantively motivated null hypothesis against an unrestricted alternative hypothesis, a Waldian t test with default priors suggested by Rouder et al. (2009) might be more appropriate.

Waldian t tests are an important example of how sequential analysis, extended to composite hypotheses, can improve statistical practice. Bayesian and frequentist statistics represent fundamentally different approaches to probability and the aim of statistical inference. Statisticians have been engaged in this debate for well over two centuries (Efron, 2005) and it seems far from over—particularly in the field of psychology, where it often resembles a theological war rather than a scientific discourse. Despite their differences, however, the approaches are "both quite legitimate" (Efron, 2005, p. 1). Therefore, rather than choosing a side (Dienes, 2011) our efforts should be focused on a reconciliation of the two schools (Berger, 2000; Little, 2006). Much like the blind men from Indostan, we might find that an integration of different approaches—rather than an ideological dispute—might bring us closer to the truth. The Waldian t test, by combining the advantages of Bayesian and frequentist procedures, is an attempt to contribute to this goal.

3 Part II: Model Families

Essentially, all models are wrong,
but some are useful.

(G. E. P. Box)

3.1 Multinomial Processing Tree Models

Psychological research relies to a large extent on simple statistical modeling such as analysis of variance or *t* tests as considered in the previous chapter. This is appropriate for tests of hypotheses on directly observable behavior or if the processes at test can be operationalized in such a way that observed responses serve as direct indicators of these processes. Many psychological theories, however, assume that observed behavior is determined by a number of qualitatively distinct, latent processes. Memory judgments, for example, might be determined by encoding, retrieval, and guessing processes (Batchelder & Riefer, 1986). Similarly, truth judgments might be the result of an interplay of knowledge and response biases (Hilbig, 2012). To formalize psychological theories about latent cognitive processes underlying observed behavior, and to allow for critical tests of their assumptions, formal mathematical measurement models constitute powerful tools (Erdfelder, Castela, Michalkiewicz, & Heck, 2015).

In 1988, Riefer and Batchelder introduced a particularly influential class of substantively motivated stochastic models for categorical data: multinomial processing tree (MPT) models. MPT models assume that observed responses stem from a finite set of discrete processing states. Specifically, MPT models explicitly specify the sequences of processing states that determine response behavior. These sequences are represented by branches in a processing tree that connects the experimental input to all possible response categories. Along these branches, the latent processing states are represented as nodes, connected by links denoting the (conditional) probabilities of entering the respective states. Based on the assumption that observed category frequencies follow a multinomial distribution, the category probabilities are modeled by these branches, thereby allowing to measure (i.e., estimate) and test the contribution of each assumed cognitive process (Erdfelder et al., 2009; Hu & Batchelder, 1994).

As an example, consider a simple hypothetical perception experiment in which participants are presented with a flash of light displayed on a screen for 100 ms in one of two temporally defined intervals. Following each of these trials, participants are asked to indicate in which of the two intervals the stimulus was presented (two-alternative forced-choice test, 2AFC). As an experimental manipulation, the stimuli are presented in two different luminous intensities, that is, high versus low intensity (see Blackwell, Pritchard, & Ohmart, 1954, for a similar experimental procedure).

Obviously, the observed performance in this task is not a process-pure measure of detection or perceptual abilities. We can assume that if a participant detects the stimulus, they will answer correctly. If they do not detect the stimulus, however, they might still give a correct response simply by guessing the interval in which the stimulus was presented. Thus, the performance measure in the 2AFC is a confound of two qualitatively different processes: a detection process representing perceptual abilities, and a guessing process.

Figure 2 displays a simple MPT model that explicitly specifies the assumed process structure underlying observed responses in the 2AFC. It is based on the two-high-threshold model of recognition memory (Snodgrass & Corwin, 1988), a prominent and well-studied example of MPT models (Bröder & Schütz, 2009). Presented with a high- or low-intensity stimulus in each trial, participants either detect the stimulus (with probabilities d_h or d_l, respectively) and answer correctly, or they do not detect the stimulus ($1 - d_h$ or $1 - d_l$). In the latter case, in a state of uncertainty, they have to guess in which interval the stimulus was presented. Their answer can either be correct (with conditional probability g) or incorrect ($1 - g$).

Within a tree, the probability of each branch leading to a response category is simply the product of (conditional) probabilities of all processing states along the branch. Across the branches, the probability of a response category is the sum of all branch probabilities ending in the respective category, thus expressing the response probabilities as a function of the model parameters (Hu & Batchelder, 1994). In our example, the probability of correctly identifying the interval in which a high-intensity stimulus was presented is thus given by

$$P(\text{Correct}|\text{High intensity}) = d_h + (1 - d_h) \cdot g. \tag{3.1}$$

There is a well-developed body of statistical theory surrounding MPT modeling, providing methods and software for maximum-likelihood as well as Bayesian parameter estimation (e.g., Heck, Arnold, & Arnold, 2018; Hu & Batchelder, 1994; Klauer, 2006, 2010; Moshagen, 2010; Singmann & Kellen, 2013; Stahl & Klauer, 2007). These methods also allow for goodness-of-fit tests to asses whether empirical data can be reconciled with the model's assumptions, and for statistical hypothesis tests on single parameters (Erdfelder et al., 2009).

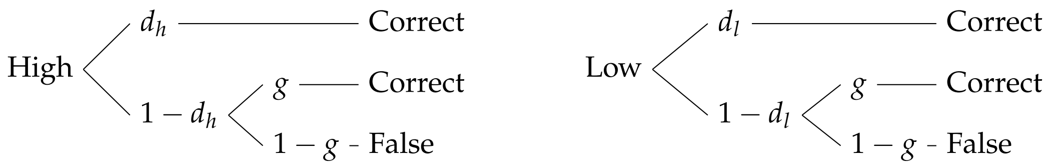

FIGURE 2: A multinomial processing tree model for a perception experiment with high versus low luminous intensity and a two-alternative forced-choice test. d_h = probability to detect the stimulus with high intensity; d_l = probability to detect the stimulus with low intensity; g = probability to guess correctly.

Beyond model fit, necessary conditions for the psychological interpretation of MPT parameter estimates are identifiability and experimental validation (Batchelder & Riefer, 1999). Identifiability denotes a one-to-one mapping of parameter values to observed data, which is necessary for the model to provide unique parameter estimates (Bamber & van Santen, 2000). In other words, an MPT model is identifiable if any set of observed, model-consistent category probabilities corresponds to one, and only one, set of parameter values. Experimental validation involves the demonstration that experimental manipulations of specific cognitive processes selectively affect the corresponding model parameters (Hilbig, 2012; Voss, Rothermund, & Voss, 2004). Only when a model has been validated, and identifiability and model fit have been established, parameter estimates allow for a psychologically meaningful interpretation as measures of latent, cognitive processes.

Over the last decades, numerous MPT models have been developed, validated, and successfully applied to substantive research questions in various branches of psychology, especially in (social-)cognitive research (see Batchelder & Riefer, 1999; Erdfelder et al., 2009; Hütter & Klauer, 2016, for reviews). MPT models have also been suggested as measurement tools for psychometric purposes. Due to the explicit formalization of latent processes underlying observed behavior and their ability to measure and disentangle them, Batchelder (1998) saw an "untapped potential" for MPT applications in individual assessments (p. 331). He coined the term *cognitive psychometrics*, repeatedly promoting the use of MPT models in clinical settings, for example, to identify specific cognitive deficits of individuals or subpopulations (Riefer, Knapp, Batchelder, Bamber, & Manifold, 2002).

In the third article of my dissertation (Schnuerch et al., 2020), we address a notable limitation of MPT model analysis which is particularly relevant in individual assessment situations or when each individual provides only a single data point (e.g., Heck, Thielmann, Moshagen, & Hilbig, 2018; Klauer, Stahl, & Erdfelder, 2007; Schild, Heck, Ścigała, & Zettler, 2019): Classical, sufficiently powered hypothesis tests typically require conservative assumptions about unknown model parameters, resulting in extremely large required sample sizes. As a remedy, we propose substantially more efficient sequential probability ratio tests for MPT models. Moreover, we consider an extension based on maximum likelihood theory which allows to test composite hypotheses without making explicit assumptions about unknown model parameters. With this approach, we show how sequential analysis can facilitate the applicability of formal measurement models such as MPT models in substantive research.

3.2 Sequential Maximum Likelihood Ratio Tests

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

Hypothesis tests for MPT parameters, that is, tests of parameter constraints representing a psychological hypothesis, typically rely on NHST (Batchelder & Riefer, 1999). To test a set of parameter constraints, we can compare the goodness-of-fit statistic of the constrained model to that of an unconstrained model. Under certain circumstances, the difference statistic follows a χ^2 distribution if the null hypothesis (i.e., the set of constraints) holds (Read & Cressie, 1988).

A sensible test requires sufficient statistical power to reject the constraints if they do, in fact, not hold in the population (Batchelder & Riefer, 1990). Power analyses for MPT models have been worked out (Erdfelder, Faul, & Buchner, 2005) and implemented in readily available software (e.g., Faul et al., 2009; Moshagen, 2010). Such an analysis, however, requires explicit assumptions about all model parameters. While test-relevant parameters are specified by the hypotheses, MPT models typically comprise additional, unknown nuisance parameters, thus rendering the hypotheses composite.

To deal with this, we can make conservative assumptions about the nuisance parameters so as to choose a sample size that ensures sufficient power for any plausible value of the unknown parameters. Such a strategy, however, may result in extremely large required sample sizes. Especially when resources are scarce, for example, in individual assessments, this is a notable limitation, either prohibiting the application of MPT models or resulting in underpowered tests and biased inference (Batchelder & Riefer, 1990).

The SPRT provides an attractive and very efficient alternative to conventional statistical tests. If all parameters are specified under the hypotheses, that is, if they are simple, the application of the SPRT to hypothesis tests in MPT models is straightforward (Schnuerch et al., 2020). It has been proven to be the optimal test for this situation and its properties (error probabilities and expected sample size as a function of true parameter values) can be determined analytically (Wald, 1947), indicating an average sample-size reduction of around 50% while holding the error rates constant. For composite hypotheses due to nuisance parameters, however, the general theory of the SPRT does not apply. This can be remedied with an extension developed by D. R. Cox (1963), which is based on asymptotic maximum likelihood (ML) theory.

Let X be a random variable, with $X \sim f(x|\theta,\, \phi)$, and assume a test of the hypothesis $\mathcal{H}_0$: $\theta = \theta_0$ versus $\mathcal{H}_1$: $\theta = \theta_1$. If ϕ is unknown, and no prior distribution can be assumed (see Schnuerch et al., 2019), D. R. Cox (1963) suggested to construct a sequential test based on

$$LR_n = \frac{f(x^n|\theta_1,\, \hat{\phi})}{f(x^n|\theta_0,\, \hat{\phi})}, \tag{3.2}$$

where $\hat{\phi}$ denotes the ML estimate of ϕ based on x^n, conditional on a model without restrictions on θ or ϕ.[3] D. R. Cox (1963) showed that if the ML estimates $\hat{\theta}$, $\hat{\phi}$ are asymptotically independent, a simple SPRT as defined in (2.5) based on (3.2) is asymptotically equivalent to that when ϕ is known. Otherwise, however, the sequential procedure must be corrected for fluctuations in the likelihood ratio caused by sampling error of $\hat{\phi}$.

For simplification, the likelihood ratio in (3.2) can be replaced by the ratio of second-order Taylor series expansions about the true parameter values θ and ϕ. As far as possible, the resulting terms are then further replaced by expressions that are asymptotically equivalent (see Breslow, 1969; D. R. Cox, 1963; Schnuerch et al., 2020, for mathematical details), finally leading to the following, simple procedure (see also Wetherill, 1975): For every integral value of n, compute the test statistic

$$T_n = n\left[\hat{\theta} - \tfrac{1}{2}(\theta_0 + \theta_1)\right], \tag{3.3}$$

with $\hat{\theta}$ denoting the ML estimate of θ based on x^n. Sampling is continued as long as

$$\frac{\mathcal{V}_{\theta\theta}}{\theta_1 - \theta_0}\log\left(\frac{\beta}{1-\alpha}\right) < T_n < \frac{\mathcal{V}_{\theta\theta}}{\theta_1 - \theta_0}\log\left(\frac{1-\beta}{\alpha}\right), \tag{3.4}$$

such that the resulting test procedure satisfies approximately the desired error rates α and β.

In Equation 3.4, $\mathcal{V}_{\theta\theta}$ denotes the $(\theta,\, \theta)$ element of the inverse of the expected Fisher information for sample size $n = 1$. This value denotes the variance of the ML estimate $\hat{\theta}$ based on a single observation, assuming observations to be independent and identically distributed (see Ly, Marsman, Verhagen, Grasman, & Wagenmakers, 2017). Thus, the sequential procedure is adjusted for the additional uncertainty about the test-relevant parameter resulting from the necessity to estimate the unknown nuisance parameter ϕ.

[3] Bartlett (1946) presented a similar idea, but his method includes separate ML estimates for ϕ under each hypothesis, conditional on $\theta = \theta_0$ and $\theta = \theta_1$, respectively.

In Schnuerch et al. (2020), we show that it is straightforward to implement this method in the context of MPT models to overcome the practical limitations described above. Since the dimensionalities of θ and ϕ do not matter (D. R. Cox, 1963), we can denote by θ the test-relevant MPT model parameters specified by the hypotheses, while ϕ denotes the vector of unknown model parameters. Thus, the test allows for hypothesis tests on specific MPT model parameters without explicit assumptions on unknown nuisance parameters.

We consider three hypothetical psychometric experiments relying on MPT models to demonstrate the SPRT and its ML extension developed by D. R. Cox (1963), which we refer to as *sequential maximum likelihood ratio test* (SMLRT). By means of simulations, we explore the core properties of the SMLRT in these examples, namely, the empirical error rates and sample-size distributions. Overall, we demonstrate that the procedure allows for decision-error probability control without assumptions about unknown parameters. At the same time, the sequential test requires notably fewer observations than classical Neyman-Pearson tests with power analyses based on the same nominal error rates and the true values of the unknown nuisance parameters.

It is important to keep in mind, however, that the test is derived from asymptotic ML theory. Thus, its assumptions may be violated for small samples, leading to premature decisions and increased error rates if the initial sample size of the sequential procedure is too small (C. P. Cox & Roseberry, 1966; Wetherill, 1975). This can be remedied by imposing a minimum number of observations that have to be sampled before making a decision. There is no definite strategy how to determine this number, however. It may vary considerably based on model complexity and effect size. Therefore, based on our simulations, we recommend an initial sample size of 25–50% of the corresponding Neyman-Pearson sample size, determined by an a priori power analysis and Monte Carlo simulations.

Another caveat concerns the assumption that observations are independent and identically distributed (i.i.d.). If contaminating effects such as fatigue, exercise, or order effects can be ruled out, this assumption is reasonable when analyzing data from one individual or when each individual provides only a single data point. For aggregate data with multiple data points from several individuals, however, the i.i.d. assumption may often be questioned (Smith & Batchelder, 2008). Ignoring the hierarchical structure and dependencies in the data may result in biased inference, particularly when the data are sampled and analyzed sequentially. Thus, the sequential approach is not suited for those cases in which there are multiple data points per individual and no reason to assume i.i.d. observations across individuals.

Especially in the context of individual assessments (i.e., cognitive psychometrics), however, the i.i.d. assumption is typically justified. At the same time, efficiency is of particular concern for reasons of limited resources. Therefore, SPRTs and SMLRTs are attractive alternatives to conventional hypothesis tests in this situation. To summarize, as we outlined theoretically and demonstrated empirically, sequential analysis can improve the applicability of MPT models considerably, thus fostering their application in research areas that could greatly benefit from this class of theoretically motivated stochastic models.

4 Part III: Application Fields

> Gedanken ohne Inhalt sind leer,
> Anschauungen ohne Begriffe sind blind.[4]
>
> *(I. Kant)*

The aim of the methods studied and developed in my dissertation is to improve statistical practice and provide psychological researchers with more reliable and efficient means to approach substantive research questions and test psychological theories. In the following sections, I describe two concrete examples from different fields of psychology where an application of the methods led to a substantial saving of required resources. In the first one, Hajnal's t test was employed to test hypotheses about the influence of age on the attraction search effect (Scharf, Fischer, & Schnuerch, 2020). The second example is an application of the SMLRT to the randomized response technique, which was used to test hypotheses about gender differences in the prevalence of casual sex.

4.1 The Attraction Search Effect

In all areas of our lives, we have to make decisions: When do we get up in the morning? What clothes do we choose to wear? Which job do we take? What name do we pick for our children? The circumstances and consequences of the decisions we make may vary greatly, but they all typically include the search for and integration of relevant information about the choice options. Whereas there is a considerable body of theory and empirical evidence on how people *integrate* information, there has been a somewhat surprising neglect of information *search* in psychological research for decades (Todd & Gigerenzer, 2012).

In the classical probabilistic inference task, participants are presented with two options (e.g., stocks) and a number of cues. Each cue provides information on each of the options (i.e., cue values), which can be either positive or negative. The cues differ in terms of validity, that is, predictive quality. The task is to choose one of the options, based on the information provided by the cues. To study information search, participants are presented with a pattern of open and concealed cue values and they can

[4] *Thoughts without contents are empty, intuitions without concepts are blind.*

choose the next cue values to open. Typically, information search is restricted to a certain number of additional openings and associated with costs.

A critical assumption of the few theoretical accounts that include information search (e.g., the "adaptive toolbox"; Gigerenzer & Todd, 1999) is that search rules are fixed (Jekel, Glöckner, & Bröder, 2018). That is, they assume that people will search for information either within cues (cue-wise) or within options (option-wise), irrespective of the information that is already available. There is a magnitude of evidence, however, that search rules vary depending on contextual features of the decision situation or the decision maker (e.g., Bröder, 2000, 2003; Glöckner & Moritz, 2008; Mata, Schooler, & Rieskamp, 2007; Mata, von Helversen, & Rieskamp, 2011; Rieskamp & Hoffrage, 2008), as well as the available information (Söllner & Bröder, 2016; Söllner, Bröder, Glöckner, & Betsch, 2014), which challenges the assumption of a fixed search rule.

An alternative, more successful account was recently presented by Jekel et al. (2018), the *integrated coherence-based decision and search model* (iCodes). The model is an extension of the parallel constraint satisfaction model for decision making (Glöckner, Hilbig, & Jekel, 2014), a network model that assumes parallel processing of all available information, represented by activation spreading through the network. Importantly, iCodes makes a new and unique prediction about information search in a decision-making situation: Assuming that, based on the available information, one of the options is more "attractive" than the other, iCodes predicts that the search is directed toward information about the more attractive option.

This *attraction search effect* (ASE) has been shown in a number of studies using classical paradigms such as hypothetical stock-market tasks (Jekel et al., 2018) or other, more realistic contexts and task formats (Scharf, Wiegelmann, & Bröder, 2019). It is assessed via the *attraction search score* (ASS). Participants are presented with an information matrix where some cue values are open while others are concealed. Their search behavior in a trial is measured by the first cue value they open, which can either be a cue value that contains information on the more attractive option (based on the initially revealed information) or one with information on the less attractive one. The ASS is calculated as the difference between the conditional probabilities of cue searches for an option when it is attractive versus when it is unattractive. An $ASS = 1$ would thus denote a perfect ASE, that is, someone always searches for information about an attractive option, but never about the unattractive one.

In Scharf et al. (2020), we investigated whether there is a *positivity effect* in the context of the ASE. The positivity effect denotes "an age-related trend that favors positive over negative stimuli in cognitive processing" (Reed & Carstensen, 2012, p. 1). Across different domains in cognitive psychology, there is consistent evidence that older adults strategically select positive information to process as a means to advance a state of emotional well-being (Mather & Carstensen, 2005). For example, in the context of decision making, older people have been shown to selectively search for positive information about the available choice options (Löckenhoff & Carstensen, 2007, 2008). In the context of the ASE, the valence of the concealed information is not known. However, as we assume one option to be more attractive based on the available information, the positivity effect implies that older adults should show a stronger tendency to search for information about this option. Hence, we expected a positivity effect on the ASE, that is, a stronger ASE among older adults than among younger adults.

Based on meta-analytic results on the positivity effect and age-related differences in cognitive processes (Mata & Nunes, 2010; Reed, Chan, & Mikels, 2014), we expected a rather small effect size (Cohen's d) under the alternative hypothesis, $\delta = 0.30$. For a one-tailed t test with $\alpha = .05$ and power $1 - \beta = .90$, an a priori power analysis revealed a required sample size of $N = 382$, assuming equal group sample sizes (Faul et al., 2009). To reduce this number, we opted for a more efficient statistical test with the same error probabilities, namely, Hajnal's t test (Schnuerch & Erdfelder, 2019).

In a preregistered quasi-experiment (http://osf.io/dy3jx), older (> 60 years) and younger adults (18–30 years) worked through 48 probabilistic inference tasks with different cue patterns and two different contexts (i.e., decision about cell-phone contracts and health-insurance providers). The patterns were constructed such that one of the options was the more attractive one (i.e., received more positive or less negative information from the cues) and additional information search was restricted. The ASS was calculated for each person across the 48 trials. We started the experiment with an initial sample of size $N = 3$, with $n_{\text{old}} = 2$ and $n_{\text{young}} = 1$, which is the minimum required sample size to estimate the standard error and calculate the t value (Schnuerch & Erdfelder, 2019). Hajnal's t test was then applied to the data after each additional observation and sampling continued as long as

$$\frac{\beta}{1-\alpha} < \frac{f(t_n | df_n,\ \delta = 0.30)}{f(t_n | df_n,\ \delta = 0)} < \frac{1-\beta}{\alpha}. \tag{4.1}$$

Figure 3 shows the development of the log likelihood ratio as the sample size increases. After 142 observations ($n_{\text{old}} = 78$, $n_{\text{young}} = 64$), the test was terminated because the upper threshold was crossed, with $LR_{142} = 19.62$. This ratio indicates that the data were almost 20 times more likely under the alternative hypothesis than under the null hypothesis. Note that Hajnal's t test is not affected by unequal sample sizes if homoscedasticity is not violated (Schnuerch & Erdfelder, 2019). In this experiment, there was no reason or indication in the data to assume a violation. Thus, we accepted the alternative hypothesis that the ASE is stronger among older adults than among younger

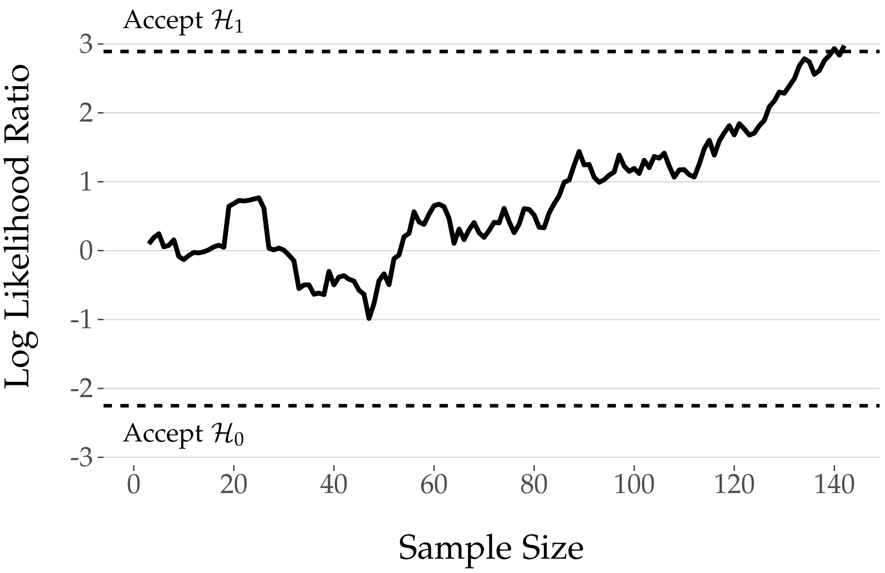

Figure 3: Development of the log likelihood ratio for Hajnal's t test. Upper and lower dashed lines represent the decision boundaries $\log(A)$ and $\log(B)$, respectively. The test terminates after $N = 142$ observations with a decision in favor of $\mathcal{H}_1$.

adults, $\hat{\delta} = 0.44$, 95% CI = [0.10, 0.78].[5]

Our study supported the hypothesis derived from the positivity effect. Importantly, it did so notably more efficiently than a classical Neyman-Pearson test. Hajnal's t test required almost 63% less observations while holding the error probabilities α and β constant. Finally, the study also supported iCodes by once again demonstrating its core prediction, that is, the ASE.

For a more critical test of the positivity effect, however, future experiments need to test the underlying mechanisms of the influence of age on the ASE. For example, is it, in fact, caused by an age-related trend to selectively process positive information (as implied by the underlying *socioemotional selectivity theory*; Carstensen, 2006)? Moreover, to exclude cohort effects as an alternative explanation, a longitudinal design is required to investigate the strength of the ASE as a function of age over the lifespan. Lastly, more research is needed on the psychological mechanisms underlying the ASE and, more generally, the processes assumed in iCodes.

[5] Note that this estimate of Cohen's d and the confidence interval are based on the assumption of a fixed sample size and not corrected for a potential bias due to the sequential analysis.

4.2 The Randomized Response Technique

Social scientists frequently rely on self-report based surveys to estimate or test the prevalence of opinions, attitudes, or behavior that cannot be observed or otherwise assessed objectively (Fox, 2016). Social surveys provide an efficient and cost-effective means to asses such attributes. They depend, however, on respondents' compliance to answer truthfully. This is a notable limitation, because respondents often have a tendency to present themselves in a positive way by responding in line with perceived expectations or social norms. Unlike random sampling errors, systematic biases such as *social desirability* corrupt the validity of prevalence estimates (Paulhus, 1991). Socially desirable responding is particularly harmful when asking about sensitive attributes, that is, opinions or behavior that violate social norms or laws, such that respondents are reluctant to disclose true answers in order to avoid negative consequences (Krumpal, 2013; Tourangeau & Yan, 2007).

As a remedy, Warner (1965) introduced a survey technique specifically designed to overcome the "evasive answer bias" (p. 63) resulting from asking sensitive or intrusive questions. Assuming that the bias is a function of perceived anonymity, Warner's technique aims to ensure the individual respondent's anonymity in order to encourage truthful responding. The basic idea of the *randomized response technique* (RRT) is that random elements enter the response process, thus stripping individual responses of all diagnostic value as to the true status of the respondent. Specifically, individuals are prompted to answer to either the sensitive statement A (e.g., *I have used illicit drugs in the past.*) or its logical opposite ¬A (*I have never used illicit drugs.*), depending on the outcome of some randomization device (e.g., the role of a die). Critically, only the respondents know the outcome and, thus, which statement they responded to. Consequently, in the RRT, an individual "Yes" or "No" response provides no valid information on whether or not the respondent possesses the sensitive attribute (e.g., consumption of illicit drugs).

When the probability p ($p \neq .50$) of the randomization device to lead to statement A is known, however, the population prevalence π of the sensitive attribute can easily be estimated from the observed responses (see Figure 4 for illustration). Let λ denote the probability of a "Yes" response, then the RRT implies that $\lambda = p\pi + (1-p)(1-\pi)$. Given the observed proportion of positive responses $\hat{\lambda}$, an unbiased estimate of π is given by

$$\hat{\pi} = \frac{\hat{\lambda} + p - 1}{2p - 1} \tag{4.2}$$

with sampling variance (in a sample of size N)

$$\mathrm{Var}(\hat{\pi}) = \frac{\pi(1-\pi)}{N} + \frac{p(1-p)}{N(2p-1)^2}. \tag{4.3}$$

The RRT has repeatedly been shown to result in higher prevalence estimates of sensitive attributes than direct questioning in the context of comparative validation studies (Horvitz, Greenberg, & Abernathy, 1976; Lensvelt-Mulders, Hox, van der Heijden, & Maas, 2005). Moreover, experimental validation studies have found that it leads to prevalence estimates that are closer to known prevalence rates (e.g., Hoffmann, Diedenhofen, Verschuere, & Musch, 2015; Moshagen, Hilbig, Erdfelder, & Moritz, 2014). Inspired by Warner's model, numerous model extensions and variations have been proposed over the last decades, aimed at improving the technique in terms of statistical properties or psychological acceptance (e.g., Clark & Desharnais, 1998; Greenberg, Abul-Ela, Simmons, & Horvitz, 1969; Kuk, 1990; Mangat, 1994; Moors, 1971; Moshagen, Musch, & Erdfelder, 2012; Yu, Tian, & Tang, 2008).

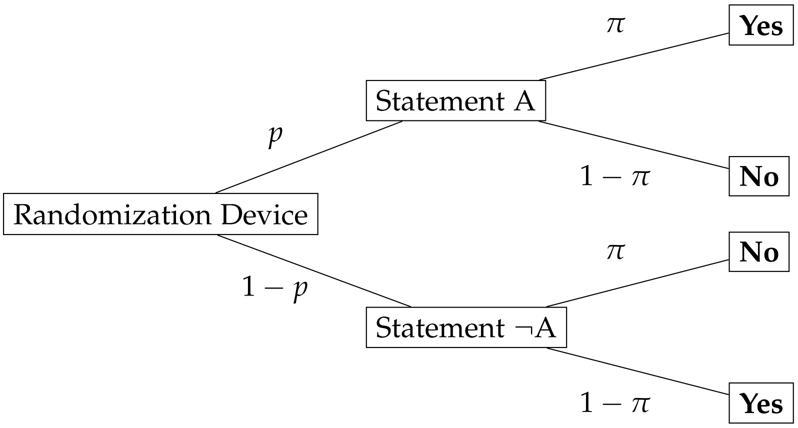

FIGURE 4: Warner's (1965) randomized response technique, with p ($p \neq .50$) denoting the known probability to receive the sensitive statement and π representing the unknown population prevalence of the sensitive attribute.

Although RRT models have successfully been applied to substantive research questions in the past (e.g., Dietz et al., 2013; Moshagen, Hilbig, & Musch, 2011; Ulrich, Pope, et al., 2018), there is a notable drawback of the RRT which may in part be responsible for the surprisingly low number of substantive applications (Blair, Imai, & Zhou, 2015): Due to the random noise that enters the response process, which is the crucial element of the RRT, the sampling variance of the estimator $\hat{\pi}$ is inflated (see Equation 4.3). This increase in variance compared to direct-questioning methods results in decreased estimation precision and low statistical power when testing hypotheses on the unknown prevalence (Ulrich, Schröter, Striegel, & Simon, 2012). Consequently, particularly when effect sizes are small or the level of anonymity is high, RRT models require extremely large sample sizes, thus rendering their application more or less infeasible for researchers with limited resources.

As a remedy, Reiber, Schnuerch, and Ulrich (2019) suggested to improve the efficiency of hypothesis tests in the RRT by means of sequential analysis. Specifically, we proposed curtailed sampling for RRT models, a simple sequential procedure for binomial tests (Wetherill, 1975). The curtailed approach requires on average fewer observations than a corresponding Neyman-Pearson test. At the same time, it is easy to apply, truncated at the Neyman-Pearson sample size, and there are closed-form solutions for unbiased estimators for the unknown prevalence (Girshick, Mosteller, & Savage, 1946).

Despite its advantages, however, curtailed sampling has a notable limitation. Like Neyman-Pearson tests, it only applies to simple hypotheses. Thus, it is not suitable for two-tailed hypothesis tests and other tests of composite hypotheses (e.g., due to nuisance parameters in multi-parameter RRT models; Clark & Desharnais, 1998; Moshagen et al., 2012). Since RRT models belong to the class of MPT models, however, a solution for this problem is given by Schnuerch et al. (2020): The SMLRT provides an efficient means to test composite hypotheses in the context of RRT models without explicit assumptions about unknown nuisance parameters. Moreover, it is straightforward to derive closed-form solutions for the test statistic and threshold values for common scenarios such as comparisons between independent prevalence estimates and multi-parameter RRT models.

In a first empirical application, this sequential RRT was used to test three competing hypotheses about gender differences in the prevalence of casual sex.[6] Grello, Welsh, and Harper (2006) define casual sex as "sexual relationships in which the partners do not define the relationship as romantic or the partner as a boyfriend or girlfriend" (p. 255). A typical finding in surveys on casual sex behavior is that incidence rates of casual sex are higher among men than among women (e.g., Herold & Mewhinney, 1993; see also Petersen & Hyde, 2010). This is somewhat surprising, as one might expect that for heterosexual men and women these numbers should match. However, whereas this reasoning is true for the total number of sexual encounters, it does not necessarily apply to incidence rates. In fact, we might consider three competing, mutually exclusive hypotheses on gender differences in the prevalence of casual sex, derived from contemporary theories on sexual behavior.

First, in a strict interpretation of the gender similarities hypothesis (Hyde, 2005), we might expect that there are no meaningful differences between genders, indicating that prevalence rates of casual sex should indeed not differ between men and women. In contrast, from the perspective of *sexual economics*, sex is considered a valuable female resource which is traded for other, non-sexual resources in a sexual marketplace (Baumeister & Vohs, 2004). According to this perspective, women as the sellers of sex should seek to maximize its value by keeping it scarce, while men should generally strive to acquire it for as low costs as possible. Thus, women should be less inclined to engage in casual sex than men, resulting in a smaller prevalence among women than

[6] I am grateful to Benjamin Hilbig for bringing my attention to this research project.

among men.

A different prediction can be derived from the evolutionary perspective of the *sexual strategies theory* (Buss & Schmitt, 1993). According to this theory, women and men pursue different strategies to increase their chances of genetic success. While men can maximize the *quantity* of offspring by maximizing the number of sexual partners, women need to maximize the *quality* of their partners to acquire the best possible genes for their offspring. Casual sex may have reproductive benefits for both genders as it increases the number of sexual partners, which benefits men, and the chance to evaluate potential long-term partners, which is important for women. However, whereas men should engage in casual sex rather indiscriminately, women should be highly selective and only engage in casual sex with those men that meet certain standards to qualify them as potential long-term partners. Thus, only the subgroup of men that possess these qualities get the opportunity for casual sex, and they should engage in casual sex with a larger group of women, resulting in a higher prevalence of casual sex among women than among men. Figure 5 illustrates the qualitative predictions of the three outlined hypotheses.

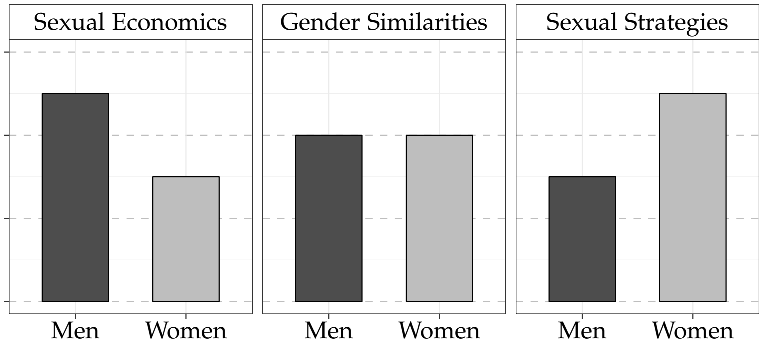

Figure 5: Qualitative predictions for casual sex prevalences among men and women.

Empirical evidence from past surveys seems to provide support for the sexual economics theory. As they relied exclusively on self-report measures, however, response biases cannot be excluded as an alternative explanation. Casual sex is considered a sensitive attribute (Baumeister & Vohs, 2004) and sexual double standards may systematically motivate men and women to over- or underreport their sexual behavior, respectively (Crawford & Popp, 2003). Therefore, to reduce social-desirability bias and critically test the three competing hypotheses, I assessed the prevalence via an RRT.

Specifically, I used the crosswise model (CWM; Yu et al., 2008), a newer model variant, which is often referred to as *non-randomized* response technique. Instead of the randomization device leading to one of two logically opposite statements, participants in the CWM are prompted to simultaneously respond to the sensitive and a neutral statement.

The task is to indicate whether the answers to the statements are identical (response 'A') or different (response 'B'). The resulting model is mathematically identical to Warner's original RRT. Let p denote the prevalence of the neutral statement, then the probability λ of an 'A' response is given by $\lambda = p\pi + (1-p)(1-\pi)$, resulting in the same estimator as for Warner's model (Equations 4.2 and 4.3).

Participants were presented with the following statements:

1. I had casual sex during the last 12 months.
2. My mother was born between January 1 and September 30.

Assuming a uniform distribution of birth rates across the months, the prevalence of the neutral statement is $p = .75$. Let π denote the prevalence of casual sex among women and $\pi + \theta$ the prevalence for men. We can then express our hypotheses in terms of θ. The prevalence π, however, is an unknown nuisance parameter. Thus, the hypotheses are composite.

Under the null hypothesis (representing the gender similarities hypothesis) I expected no difference between the prevalence rates, $\theta_0 = 0$. The sexual economics and sexual strategies theories both predicted a difference, but in opposite directions. Thus, the alternative hypothesis is the two-tailed hypothesis $\theta_1 \neq 0$, with $\theta_1 < 0$ representing the sexual strategies account and $\theta_1 > 0$ denoting the sexual economics approach. Following previous results from direct-questioning studies (Grello et al., 2006), I specified an effect size of $|\theta_1| = .15$ under the alternative hypothesis.

An a priori power analysis requires an assumption about the nuisance parameter π. To ensure a sufficiently powered test for the above hypotheses, the analysis was based on the conservative assumption $\pi = .50$ and $\alpha = \beta = .05$, resulting in a required sample size of $N = 2,292$, assuming group sample sizes to be equal (Fleiss, Levin, & Paik, 2003; Ulrich et al., 2012). Instead, I used the SMLRT without any assumption on π. Note that although I am testing hypotheses on θ (i.e., the difference between prevalence rates), the test is based on the linear transformation λ of these prevalence rates. Let λ_i^θ denote the predicted difference in λ corresponding to hypothesis i, $i = 0, 1$. Then, in a straightforward application of Equations 3.3 and 3.4, the test statistic computed after every single additional observation is given by

$$T_n = \frac{n_1 + n_2}{2} \left(\frac{n_1 x_1 - n_2 x_2}{n_1 n_2} - \frac{\lambda_0^\theta + \lambda_1^\theta}{2} \right), \tag{4.4}$$

with n_1, n_2 denoting the sample sizes in each group and x_1, x_2 denoting the observed number of 'A' responses in each group. Note that although it allows for unequal group sizes, the derivation of this formula is based on the assumption that observations are sampled from both populations with equal probabilities. Correspondingly, the term

$\mathcal{V}_{\lambda^\theta \lambda^\theta}$ for calculating the threshold values is given by

$$\frac{(n_1 + n_2)\left[n_1^3 x_2 (n_2 - x_2) + n_2^3 x_1 (n_1 - x_1)\right]}{2 n_1^3 n_2^3}. \tag{4.5}$$

To extend the SMLRT to the case of two-tailed hypotheses, I superimposed two one-tailed tests with Type-I error probabilities $\alpha/2$ (Armitage, 1950). The procedure continues until one of the tests crosses the upper threshold or, alternatively, both tests have crossed the lower threshold. In the former case, the alternative hypothesis of the test that crossed the threshold is accepted, while in the latter case, the common null hypothesis is accepted. Following the recommendations given in Schnuerch et al. (2020), I performed Monte Carlo simulations prior to the study to define a lower boundary for the sample size of the SMLRT. Based on these simulations, I defined the initial sample size $N_{\text{min}} = 20$.

The study was conducted online. All participants received detailed instructions on the topic and the questioning technique. Participants who were unable to answer control questions about the instructions were excluded. After the indirect question, participants disclosed demographic information. Only those who identified as heterosexual men and women were included in the analyses.

Figure 6 depicts the development of the test statistics as a function of the sample size. After 40 observations, the test of the sexual strategies hypothesis accepted the null hypothesis. The other test continued until it also reached the lower boundary. Thus, after 233 observations, the test was terminated and the null hypothesis was accepted.

At the point of termination, the sample comprised 81 male and 152 female participants. The prevalence estimates in the two groups were $\hat{\pi}_{\text{male}} = .41$ and $\hat{\pi}_{\text{female}} = .31$, resulting in the estimate $\hat{\theta} = .10$ with the 95% adjusted Wald CI $[0, .24]$ (Agresti & Coull, 1998). Note, however, that these estimates are not corrected for a potential bias due to sequential analysis.

To summarize, the study supports the strict interpretation of the gender similarities hypothesis that there is no meaningful difference in the prevalence of casual sex between men and women. What is more, because the test was based on the SMLRT, I was able to accept the null hypothesis already after 233 observations. Compared with a classical Neyman-Pearson test with the same error probabilities, this is a reduction of almost 90%. Thus, this study nicely demonstrates the benefits of sequential analysis by improving the efficiency and, thus, facilitating the application of RRT models in substantive research.

A potential limitation of the study is the considerable difference in group sample sizes. This might have been the result of sampling error, in which case the test's properties are unaffected. If, however, observations were not sampled with equal probabilities

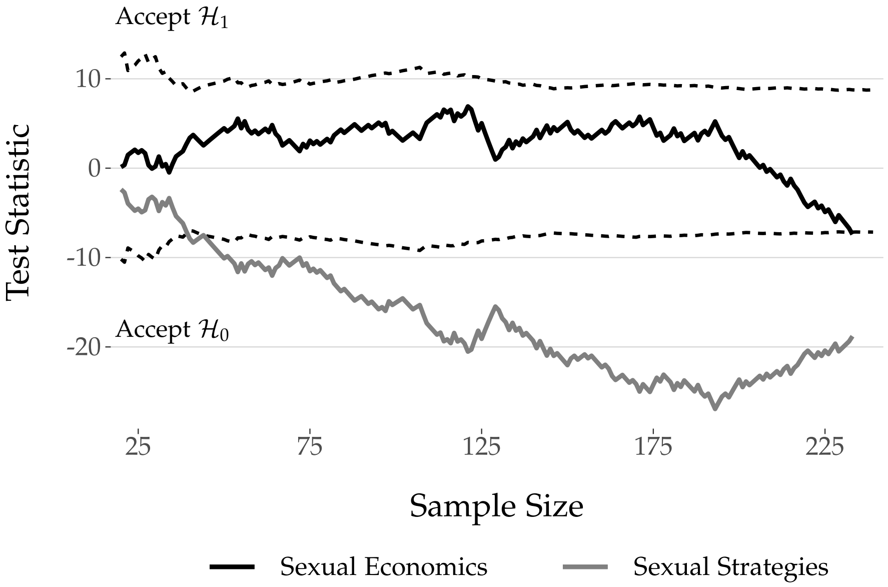

Figure 6: Development of the test statistics. Upper and lower dashed lines represent the upper and lower threshold, respectively. Sampling terminated after $N = 233$ observations.

from both groups, error rates of the procedure might be increased. Simulations indicate that, for the observed ratio of group sizes, the reduction in statistical power is negligible. Nevertheless, future studies should replicate the finding with constant sampling probabilities for both groups.

5 Discussion

In my dissertation, I implemented, further developed, and examined the properties of sequential hypothesis tests. Specifically, I studied three methods to extend the most efficient sequential test, the sequential probability ratio test (Wald, 1945, 1947), to the case of composite hypotheses.

With few exceptions, sequential hypothesis tests have largely been ignored in psychological research (Botella et al., 2006; Lang, 2017). This was partly due to the mathematical complexity of the procedures, especially before the emergence of high-performance computers. What is more, the procedures were typically designed for simple hypotheses. In practical applications, however, hypotheses are rarely simple, which limits the usefulness of these test procedures considerably.

Throughout the literature, different solutions have been developed to extend sequential tests to the common case of composite hypotheses (Wetherill, 1975). By seizing on and implementing these methods, I demonstrate that the scope of sequential hypothesis tests is much wider than typically assumed and that they are, by no means, limited to simple hypotheses. What is more, I aim to show how SPRTs can be used to overcome a number of pervasive problems that typically arise in the context of statistical hypothesis testing. Conventional procedures that control error probabilities often require extremely large sample sizes to ensure a sufficiently powered test. They do not allow for distributional hypotheses and require explicit assumptions about all parameters, even unknown nuisance parameters. SPRTs, in contrast, can be extended to handle distributional hypotheses and unknown parameters, and they are on average much more efficient than conventional tests. By implementing and examining sequential procedures for composite hypotheses, I hope to contribute to the improvement of statistical practice in psychology, such that substantive researchers are equipped with more efficient and more reliable means to formalize theories and critically test hypotheses.

In the first article (Schnuerch & Erdfelder, 2019), we show how the SPRT is easily extended to the common *t*-test situation, based on methods by D. R. Cox (1952b), Rushton (1950), and Hajnal (1961). Sample-size requirements of conventional tests very often result in underpowered experiments that are prone to producing unreliable results (Psychonomic Society, 2019). SPRT *t* tests control error probabilities in the same manner as Neyman-Pearson tests, and much more reliably so than Bayesian *t* tests. At the same time, they require on average substantially smaller sample sizes. Considering the amount of resources societies spend on research every single year, researchers have an ethical obligation to use these resources to the best of their possibilities (Lakens, 2014). The SPRT, as we demonstrate in the article, does so while controlling statistical decision errors.

In the second article (Schnuerch et al., 2019), we use the SPRT to bridge the gap between frequentist and Bayesian hypothesis tests. By conceptualizing the Bayes factor as a simple likelihood ratio for distributional hypotheses, as represented by the prior distributions, and combining it with decision thresholds of the SPRT, we develop a sequential design for Bayesian *t* tests that combines the advantages of both statistical worlds: The Waldian *t* test allows for statistical decisions with controlled error probabilities conditional on the hypotheses tested. At the same time, since the Bayes factor remains unchanged, it preserves the fully Bayesian justification and interpretation as a measure of evidence and updating factor of subjective belief. Thus, the SPRT provides a useful means to reconcile the somewhat ideological quarrel between frequentist and Bayesian statistical methods.

Apart from its potential role in the Bayesian-frequentist debate, the Waldian *t* test is an important addition to Hajnal's *t* test considered in the first article. Wheres the latter is more appropriate when a point alternative is specified (e.g., representing a minimum relevant effect size) for which a test with some upper-bound error probability is required, the former is more appropriate when a substantively motivated null hypothesis is tested against an unrestricted alternative hypothesis. In this case, the prior distribution of the Bayes factor represents a weight function for plausible, non-zero effect sizes under the alternative hypothesis. The resulting Waldian *t* test allows for a test of this distributional hypothesis against a point null hypothesis with controlled error probabilities.

In the last article (Schnuerch et al., 2020), we seize on yet another method to extend the SPRT to the case of composite hypotheses to improve the applicability of formal measurement models in substantive research. Hypothesis tests in multinomial processing tree models very often require conservative assumptions about nuisance parameters and extremely large samples to ensure sufficient statistical power (Batchelder & Riefer, 1990). This may limit their applicability in situations with scarce resources, for example, individual assessments. We show how a method developed by D. R. Cox (1963) can be applied to parameter tests in MPT models and how this sequential maximum likelihood ratio test increases the efficiency of those tests, while controlling decision error probabilities *without* explicit assumptions about unknown nuisance parameters.

To summarize, in the three articles of my dissertation I studied three different methods to extend the SPRT to the case of composite hypotheses. I demonstrated the beneficial properties of these methods and how they can be used to remedy practical problems of statistical hypothesis testing. For illustration, I reported two concrete examples where the methods were applied to substantive research questions. Using Hajnal's *t* test, we tested a hypothesis on the influence of age on the attraction search effect almost 63% more efficiently than with a corresponding Neyman-Pearson test (Scharf et al., 2020). In the second example, to test three competing hypotheses on gender differences in the prevalence of casual sex, I implemented the SMLRT in the context of the randomized response technique. The sequential procedure allowed to test the composite hypotheses without assumptions about unknown nuisance parameters. Moreover, the test required almost 90% fewer observations than a Neyman-Pearson test.

By explicitly locating the projects of my dissertation in SMiP's cuboid model, I also hope to demonstrate the benefits of a holistic perspective on the research process. It is important to consider all dimensions: the models that are formalized, statistical instantiations of psychological theories; the statistical techniques employed to estimate and test hypotheses on parameters of these models; and the substantive fields that motivate the research questions, theories, and hypotheses. Only by looking at all dimensions in interaction, the process can evolve as a whole and produce the scientific progress that justifies the resources societies spend on it.

5.1 Limitations and Future Directions

Throughout this thesis and the projects reported herein, we assumed that, when applying the sequential procedures, a decision will eventually be made. Although the SPRT has been proven to be a closed test (Wald, 1947, Appendix A.1), this proof is of theoretical interest only. In practical applications, it cannot be guaranteed that the test continues until a threshold is reached because there is no definite upper limit to the sample size. Thus, it may well happen that the sampling process is terminated prematurely due to practical constraints such as limited time or financial resources. Although the risk for extremely large samples is small (Schnuerch & Erdfelder, 2019; Schnuerch et al., 2020; Schnuerch et al., 2019), there is no guarantee that a non-truncated sequential procedure terminates with a reasonably small sample size.

If a statistical decision is made upon premature termination, the error probabilities of the procedure are no longer controlled (Schnuerch & Erdfelder, 2019). Therefore, a practical remedy in such a case would be to terminate sampling without a decision. In case of Hajnal's *t* test and the Waldian *t* test, the likelihood ratio at this point could still be interpreted as a continuous measure of evidence (Royall, 1997). If a decision is required, however, this constitutes a rather unsatisfying solution. Moreover, in the SMLRT, the test statistic does not have an evidential interpretation.

This latent risk to end up with extremely large samples, however unlikely, is another factor that might have prevented non-truncated sequential procedures from more widespread application in substantive research. Truncated procedures such as group sequential tests (Proschan et al., 2006) or curtailed sampling (Reiber et al., 2019) do not have this risk. However, the truncation comes at the cost of average efficiency (Schnuerch & Erdfelder, 2019). Moreover, they are limited to simple hypotheses. Armitage (1957) developed a class of "restricted sequential procedures" (p. 10). However, these procedures are not optimal, either, and rather limited in scope (Wetherill, 1975). Therefore, future research is needed to investigate possibilities to truncate the SPRT and its extensions at some predefined sample size $N_{\max}$ without compromising error rates and efficiency.

A second limitation of the sequential procedures reported herein concerns effect-size estimation. Hypothesis testing is an integral part of the scientific process (Morey, Rouder, Verhagen, & Wagenmakers, 2014). Therefore, sequential procedures aiming to increase the efficiency of hypothesis tests are an important addition to the psychological researcher's toolkit. Apart from hypothesis testing, however, precise and unbiased estimation of the effect size is frequently required. Unbiased estimation following fixed-sample tests is typically straightforward. Following a sequential test, in contrast, the distribution of conventional estimators is often distorted considerably, with small samples systematically overestimating and large samples systematically underestimating the true effect size (Whitehead, 1986). Since the sample-size distribution is not symmetric, either, effect-size estimation following sequential procedures is typically biased (e.g., Emerson & Fleming, 1990; Fan, DeMets, & Lan, 2004; Goodman, 2007; Mueller, Montori, Bassler, Koenig, & Guyatt, 2007; Schönbrodt & Wagenmakers, 2018; Stallard, Todd, & Whitehead, 2008; Zhang et al., 2012).

For certain sequential testing problems, unbiased estimators have been derived. Girshick et al. (1946) presented a number of theorems that gave rise to unbiased estimators for a range of sequential binomial tests, among others for curtailed sampling plans (see also Reiber et al., 2019) and the SPRT. In the latter case, the estimator derived by Girshick et al. (1946) is limited to the one-tailed one-sample binomial test. Moreover, unless the sample sizes are very small, the exact estimator is computationally intractable and, thus, of little practical use. D. R. Cox (1952a) derived a simple approximation, however, which appears to be close to the exact estimator.

For many other, more complex situations, there are no exact unbiased estimators following a sequential test. A closer analysis of this issue reveals, however, that the drawback is not as severe as it may seem (Goodman, 2007): A bias resulting from exclusively taking into account those studies that accepted the alternative hypothesis is not surprising, as it is based on a loss of information, not the sequential procedure itself. In fact, it is comparable with the over-estimation of effect sizes resulting from publication bias (Ulrich, Miller, & Erdfelder, 2018). Aggregating across all studies, irrespective of the point of termination, reduces the bias considerably (Schönbrodt & Wagenmakers, 2018). What is more, meta-analytic effect-size estimates that take into account not only all studies but also the sample size (i.e., precision) underlying each estimate are, in fact, unbiased (see Schönbrodt et al., 2017).

In my thesis, I considered the methods for extending the SPRT to the case of composite hypotheses in specific situations (e.g., t tests, MPT models). They are, however, much more general. Thus, the approaches presented in my thesis should be extended to other situations which are relevant for psychological researchers. For example, D. R. Cox's theorem (1952b) does not only apply to t tests but also to analysis of variance (F tests), χ^2 tests, and tests of correlation coefficients. The SMLRT could be applied to more complex models such as multiple regression models or other classes of cognitive models besides MPT models. Implementing and examining the properties of SPRTs for other scenarios would substantially increase the number of research questions to which the tests can be applied, thus making the benefits of sequential analysis available to a broader range of psychological researchers.

In the same vein, the general framework underlying Waldian t tests applies to any Bayes factor with proper prior distributions. Thus, an extension to Bayes factors with different priors than those suggested by Rouder et al. (2009), as well as to Bayes factors for other experimental designs (e.g., analysis of variance; Rouder, Morey, Speckman, & Province, 2012) is straightforward. Less straightforward, but equally relevant, would be an extension to other Bayes factor concepts, for example, adjusted fractional Bayes factors, which are based on implicit priors specified by a fraction of the data (Gu, Mulder, & Hoijtink, 2018; Hoijtink, Mulder, van Lissa, & Gu, 2019).

The main purpose of the methods presented in this thesis is the application to substantive research questions. The methods can only improve statistical practice and psychological research if they are put to use. In Chapter 4, I already presented two empirical applications that nicely demonstrated the advantages of sequential techniques in substantive research (e.g., Scharf et al., 2020).

One area in which sequential hypothesis tests might prove particularly useful in the future are replications (Lakens, 2014). The current crisis in psychology has made unequivocally clear the need for sufficiently powered replications (Asendorpf et al., 2013). This puts an additional strain on the available resources, thus increasing the need for efficient statistical methods. At the same time, replication studies typically have quite specific expectations about the effect size (e.g., assuming the lower limit of the 80% or 95% CI of the original effect-size estimate; Perugini, Gallucci, & Costantini, 2014). Consequently, replication studies represent an ideal situation for the application of sequential methods as presented in this thesis. By increasing the number of efficient, sufficiently-powered replication studies, sequential hypothesis tests could greatly benefit the entire field of psychology and play an important part in improving the replicability of psychological research.

5.2 Conclusion

Considering the amount of resources dedicated to science, researchers have an ethical and a societal obligation to use these resources in the best possible way. In my thesis, I promote a class of particularly efficient statistical techniques that can help fulfill this obligation: sequential analysis. By implementing, further developing, and examining extensions of the sequential probability ratio test to composite hypotheses, I show how these methods can be used in psychological research. They test hypotheses with nearly optimal efficiency, reliably control decision-error probabilities, unify the advantages of frequentist and Bayesian methods, and improve the applicability of stochastic measurement models. Thus, by addressing and overcoming a number of practical problems of statistical hypothesis testing, the sequential procedures presented herein have the potential to sustainably improve statistical practice in psychological research.

6 Bibliography

Agresti, A., & Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, *52*, 119–126. doi:10.1080/00031305.1998.10480550

Armitage, P. (1947). Some sequential tests of student's hypothesis. *Supplement to the Journal of the Royal Statistical Society*, *9*, 250–263. doi:10.2307/2984117

Armitage, P. (1950). Sequential analysis with more than two alternative hypotheses, and its relation to discriminant function analysis. *Journal of the Royal Statistical Society. Series B (Methodological)*, *12*, 137–144. doi:10.1111/j.2517-6161.1950.tb00050.x

Armitage, P. (1957). Restricted sequential procedures. *Biometrika*, *44*, 9–26. doi:10.2307/2333237

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., ... Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*, 108–119. doi:10.1002/per.1919

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

Bamber, D., & van Santen, J. P. H. (2000). How to assess a model's testability and identifiability. *Journal of Mathematical Psychology*, *44*, 20–40. doi:10.1006jmps.1999.1275

Barnard, G. A. (1947). The meaning of a significance level. *Biometrika*, *34*, 179–182. doi:10.2307/2332521

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

Bartlett, M. S. (1946). The large-sample theory of sequential tests. *Mathematical Proceedings of the Cambridge Philosophical Society*, *42*, 239–244. doi:10.1017/S0305004100022994

Batchelder, W. H. (1998). Multinomial processing tree models and psychological assessment. *Psychological Assessment*, *10*, 331–344. doi:10.1037/1040-3590.10.4.331

Batchelder, W. H., & Riefer, D. M. (1986). The statistical analysis of a model for storage and retrieval processes in human memory. *British Journal of Mathematical and Statistical Psychology*, *39*, 129–149. doi:10.1111/j.2044-8317.1986.tb00852.x

Batchelder, W. H., & Riefer, D. M. (1990). Multinomial processing models of source monitoring. *Psychological Review*, *97*, 548–564. doi:10.1037/0033-295X.97.4.548

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review*, *6*, 57–86. doi:10.3758/BF03210812

Baumeister, R. F., & Vohs, K. D. (2004). Sexual economics: Sex as female resource for social exchange in heterosexual interactions. *Personality and Social Psychology Review*, *8*, 339–363. doi:10.1207/s15327957pspr0804_2

Begley, C. G., & Ioannidis, J. P. A. (2015). Reproducibility in science: Improving the standard for basic and preclinical research. *Circulation Research*, *116*, 116–126. doi:10.1161/CIRCRESAHA.114.303819

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., ... Camerer, C. (2018). Redefine statistical significance. *Nature Human Behaviour*, *2*, 6–10. doi:10.1038/s41562-017-0189-z

Berger, J. O. (2000). Bayesian analysis: A look at today and thoughts of tomorrow. *Journal of the American Statistical Association*, *95*, 1269–1276. doi:10.1080/01621459.2000.10474328

Berger, J. O. (2003). Could Fisher, Jeffreys and Neyman have agreed on testing? *Statistical Science*, *18*, 1–32. doi: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* (Vol. 1, pp. 378–386). Hoboken, NJ: Wiley.

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

Blackwell, H. R., Pritchard, B. S., & Ohmart, J. G. (1954). Automatic apparatus for stimulus presentation and recording in visual threshold experiments. *Journal of the Optical Society of America*, *44*, 322–326. doi:10.1364/JOSA.44.000322

Blair, G., Imai, K., & Zhou, Y.-Y. (2015). Design and analysis of the randomized response technique. *Journal of the American Statistical Association*, *110*, 1304–1319. doi:10.1080/01621459.2015.1050028

Botella, J., Ximénez, C., Revuelta, J., & Suero, M. (2006). Optimization of sample size in controlled experiments: The CLAST rule. *Behavior Research Methods*, *38*, 65–76. doi:10.3758/BF03192751

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

Breslow, N. (1969). On large sample sequential analysis with applications to survivorship data. *Journal of Applied Probability*, *6*, 261–274. doi:10.2307/3211997

Bröder, A. (2000). Assessing the empirical validity of the "Take-the-best" heuristic as a model of human probabilistic inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*, 1332–1346. doi:10.1037//0278-7393.26.5.1332

Bröder, A. (2003). Decision making with the "adaptive toolbox": Influence of environmental structure, intelligence, and working memory load. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 611–625. doi:10.1037/0278-7393.29.4.611

Bröder, A., & Schütz, J. (2009). Recognition ROCs are curvilinear—or are they? On premature arguments against the two-high-threshold model of recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*, 587–606. doi:10.1037/a0015279

Bundesministerium für Bildung und Forschung. (2019). *Bildung und Forschung in Zahlen 2019 [Education and research in figures 2019]*. Retrieved from https://www.datenportal.bmbf.de/portal/de/B1.html

Buss, D. M., & Schmitt, D. P. (1993). Sexual strategies theory: An evolutionary perspective on human mating. *Psychological Review*, *100*, 204–232. doi:10.4324/9781351153683

Carstensen, L. L. (2006). The influence of a sense of time on human development. *Science*, *312*, 1913–1915. doi:10.1126/science.1127488

Chalmers, I., Bracken, M. B., Djulbegovic, B., Garattini, S., Grant, J., Gülmezoglu, A. M., ... Oliver, S. (2014). How to increase value and reduce waste when research priorities are set. *The Lancet*, *383*, 156–165. doi:10.1016/S0140-6736(13)62229-1

Chalmers, I., & Glasziou, P. (2009). Avoidable waste in the production and reporting of research evidence. *The Lancet*, *374*, 86–89. doi:10.1016/S0140-6736(09)60329-9

Clark, S. J., & Desharnais, R. A. (1998). Honest answers to embarrassing questions: Detecting cheating in the randomized response model. *Psychological Methods*, *3*, 160–168. doi:10.1037/1082-989X.3.2.160

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159. doi:10.1037/0033-2909.112.1.155

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

Cox, C. P., & Roseberry, T. D. (1966). A large sample sequential test, using concomitant information, for discrimination between two composite hypotheses. *Journal of the American Statistical Association*, *61*, 357–367. doi:10.2307/2282824

Cox, D. R. (1952a). A note on the sequential estimation of means. *Mathematical Proceedings of the Cambridge Philosophical Society*, *48*, 447–450. doi:10 . 1017 / S0305004100027857

Cox, D. R. (1952b). Sequential tests for composite hypotheses. *Mathematical Proceedings of the Cambridge Philosophical Society*, *48*, 290–299. doi:10.1017/S030500410002764X

Cox, D. R. (1963). Large sample sequential tests for composite hypotheses. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, *25*, 5–12.

Crawford, M., & Popp, D. (2003). Sexual double standards: A review and methodological critique of two decades of research. *Journal of Sex Research*, *40*, 13–26. doi:10.1080/00224490309552163

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

Deutsche Forschungsgemeinschaft. (2019). *Jahresbericht 2018 [Annual report 2018]*. Retrieved from https://www.dfg.de/dfg_profil/jahresbericht/index.html

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

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

Dietz, P., Striegel, H., Franke, A. G., Lieb, K., Simon, P., & Ulrich, R. (2013). Randomized response estimates for the 12-month prevalence of cognitive-enhancing drug use in university students. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy*, *33*, 44–50. doi:10.1002/phar.1166

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, *6*, 621. doi:10.3389/fpsyg.2015.00621

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

Efron, B. (2005). Bayesians, frequentists, and scientists. *Journal of the American Statistical Association*, *100*, 1–5. doi:10.1198/016214505000000033

Emerson, S. S., & Fleming, T. R. (1990). Parameter estimation following group sequential hypothesis testing. *Biometrika*, *77*, 875–892. doi:10.2307/2337110

Erdfelder, E., Auer, T.-S., Hilbig, B. E., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial processing tree models: A review of the literature. *Zeitschrift für Psychologie / Journal of Psychology*, *217*, 108–124. doi:10.1027/0044-3409.217.3.108

Erdfelder, E., Castela, M., Michalkiewicz, M., & Heck, D. W. (2015). The advantages of model fitting compared to model simulation in research on preference construction. *Frontiers in Psychology*, *6*, 140. doi:10.3389/fpsyg.2015.00140

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

Erdfelder, E., Faul, F., & Buchner, A. (2005). Power analysis for categorical methods. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (pp. 1565–1570). Chichester, UK: Wiley.

Etz, A., & Vandekerckhove, J. (2018). Introduction to Bayesian inference for psychology. *Psychonomic Bulletin & Review*, *25*, 5–34. doi:10.3758/s13423-017-1262-3

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

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149–1160. doi:10.3758/BRM.41.4.1149

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *222*, 309–368. doi:10.1098/rsta.1922.0009

Fisher, R. A. (1935a). *The design of experiments*. Edinburgh: Oliver & Boyd.

Fisher, R. A. (1935b). The fiducial argument in statistical inference. *Annals of Eugenics*, *6*, 391–398. doi:10.1111/j.1469-1809.1935.tb02120.x

Fisher, R. A. (1955). Statistical Methods and Scientific Induction. *Journal of the Royal Statistical Society: Series B (Methodological)*, *17*, 69–78. doi:10.1111/j.2517-6161.1955.tb00180.x

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical methods for rates and proportions* (3rd ed). Hoboken, NJ: Wiley.

Fox, J. A. (2016). *Randomized response and related methods for surveying sensitive data* (2nd ed.). Los Angeles: SAGE Publications, Inc.

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. (1998). We need statistical thinking, not statistical rituals. *Behavioral and Brain Sciences*, *21*, 199–200. doi:10.1017/S0140525X98281167

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

Gigerenzer, G., & Todd, P. M. (1999). Fast and frugal heuristics: The adaptive toolbox. In G. Gigerenzer, P. M. Todd, & ABC Research Group (Eds.), *Simple heuristics that make us smart* (pp. 3–34). Oxford: Oxford University Press.

Girshick, M. A., Mosteller, F., & Savage, L. J. (1946). Unbiased estimates for certain binomial sampling problems with applications. *The Annals of Mathematical Statistics*, *17*, 13–23.

Glöckner, A., Hilbig, B. E., & Jekel, M. (2014). What is adaptive about adaptive decision making? A parallel constraint satisfaction account. *Cognition*, *133*, 641–666. doi:10.1016/j.cognition.2014.08.017

Glöckner, A., & Moritz, S. (2008). A fine-grained analysis of the jumping to conclusions bias in schizophrenia: Data-gathering, response confidence, and information integration. *Judgment and Decision Making*, *4*, 587–600. doi:10.2139/ssrn.1313623

Goodman, S. N. (1993). P values, hypothesis tests, and likelihood: Implications for epidemiology of a neglected historical debate. *American Journal of Epidemiology*, *137*, 485–496. doi:10.1093/oxfordjournals.aje.a116700

Goodman, S. N. (2007). Stopping at nothing? Some dilemmas of data monitoring in clinical trials. *Annals of Internal Medicine*, *146*, 882–887. doi:10.7326/0003-4819-146-12-200706190-00010

Greenberg, B. G., Abul-Ela, A.-L. A., Simmons, W. R., & Horvitz, D. G. (1969). The unrelated question randomized response model: Theoretical framework. *Journal of the American Statistical Association*, *64*, 520–539. doi:10.1080/01621459.1969.10500991

Grello, C. M., Welsh, D. P., & Harper, M. S. (2006). No strings attached: The nature of casual sex in college students. *Journal of Sex Research*, *43*, 255–267. doi:10.1080/00224490609552324

Gu, X., Mulder, J., & Hoijtink, H. (2018). Approximated adjusted fractional Bayes factors: A general method for testing informative hypotheses. *British Journal of Mathematical and Statistical Psychology*, *71*, 229–261. doi:10.1111/bmsp.12110

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

Heck, D. W., Arnold, N. R., & Arnold, D. (2018). TreeBUGS: An R package for hierarchical multinomial-processing-tree modeling. *Behavior Research Methods*, *50*, 264–284. doi:10.3758/s13428-017-0869-7

Heck, D. W., Thielmann, I., Moshagen, M., & Hilbig, B. E. (2018). Who lies? A large-scale reanalysis linking basic personality traits to unethical decision making. *Judgment and Decision Making*, *13*, 356–371.

Hempel, C. G., & Oppenheim, P. (1948). Studies in the logic of explanation. *Philosophy of Science*, *15*, 135–175. doi:10.1086/286983

Herold, E. S., & Mewhinney, D.-a. K. (1993). Gender differences in casual sex and AIDS prevention: A survey of dating bars. *Journal of Sex Research*, *30*, 36–42. doi:10.1080/00224499309551676

Hilbig, B. E. (2012). How framing statistical statements affects subjective veracity: Validation and application of a multinomial model for judgments of truth. *Cognition*, *125*, 37–48. doi:10.1016/j.cognition.2012.06.009

Hoffmann, A., Diedenhofen, B., Verschuere, B., & Musch, J. (2015). A strong validation of the crosswise model using experimentally-induced cheating behavior. *Experimental Psychology*, *62*, 403–414. doi:10.1027/1618-3169/a000304

Hoijtink, H., Mulder, J., van Lissa, C., & Gu, X. (2019). A tutorial on testing hypotheses using the Bayes factor. *Psychological Methods*. Advance online publication. doi:10.1037/met0000201

Horvitz, D. G., Greenberg, B. G., & Abernathy, J. R. (1976). Randomized response: A data-gathering device for sensitive questions. *International Statistical Review / Revue Internationale de Statistique*, *44*, 181–196. doi:10.2307/1403276

Hu, X., & Batchelder, W. H. (1994). The statistical analysis of general processing tree models with the EM algorithm. *Psychometrika*, *59*, 21–47. doi:10.1007/BF02294263

Hütter, M., & Klauer, K. C. (2016). Applying processing trees in social psychology. *European Review of Social Psychology*, *27*, 116–159. doi:10.1080/10463283.2016.1212966

Hyde, J. S. (2005). The gender similarities hypothesis. *American Psychologist*, *60*, 581–592. doi:10.1037/0003-066X.60.6.581

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

Ioannidis, J. P. A., Greenland, S., Hlatky, M. A., Khoury, M. J., Macleod, M. R., Moher, D., . . . Tibshirani, R. (2014). Increasing value and reducing waste in research design, conduct, and analysis. *The Lancet*, *383*, 166–175. doi:10.1016/S0140-6736(13)62227-8

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

Jekel, M., Glöckner, A., & Bröder, A. (2018). A new and unique prediction for cue-search in a parallel-constraint satisfaction network model: The attraction search effect. *Psychological Review*, *125*, 744–768. doi:10.1037/rev0000107

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

Kalbfleisch, J. G., & Sprott, D. A. (1967). Fiducial probability. *Statistische Hefte*, *8*, 99–109. doi:10.1007/BF02923493

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

Kendall, M. G., & Stuart, A. (1969). *The advanced theory of statistics* (3rd ed). London: Griffin.

Klauer, K. C. (2006). Hierarchical multinomial processing tree models: A latent-class approach. *Psychometrika, 71*, 7–31. doi:10.1007/s11336-004-1188-3

Klauer, K. C. (2010). Hierarchical multinomial processing tree models: A latent-trait approach. *Psychometrika, 75*, 70–98. doi:10.1007/s11336-009-9141-0

Klauer, K. C., Stahl, C., & Erdfelder, E. (2007). The abstract selection task: New data and an almost comprehensive model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*, 680–703. doi:10.1037/0278-7393.33.4.680

Krumpal, I. (2013). Determinants of social desirability bias in sensitive surveys: A literature review. *Quality & Quantity, 47*, 2025–2047. doi:10.1007/s11135-011-9640-9

Kuk, A. Y. C. (1990). Asking sensitive questions indirectly. *Biometrika, 77*, 436–438. doi:10.2307/2336828

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

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

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

Lensvelt-Mulders, G. J. L. M., Hox, J. J., van der Heijden, P. G. M., & Maas, C. J. M. (2005). Meta-analysis of randomized response research: Thirty-five years of validation. *Sociological Methods & Research, 33*, 319–348. doi:10.1177/0049124104268664

Little, R. J. (2006). Calibrated Bayes: A Bayes/Frequentist roadmap. *The American Statistician, 60*, 213–223. doi:10.1198/000313006X117837

Löckenhoff, C. E., & Carstensen, L. L. (2007). Aging, emotion, and health-related decision strategies: Motivational manipulations can reduce age differences. *Psychology and Aging, 22*, 134–146. doi:10.1037/0882-7974.22.1.134

Löckenhoff, C. E., & Carstensen, L. L. (2008). Decision strategies in health care choices for self and others: Older but not younger adults make adjustments for the age of the decision target. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences, 63*, P106–P109. doi:10.1093/geronb/63.2.P106

Ly, A., Marsman, M., Verhagen, J., Grasman, R. P., & Wagenmakers, E.-J. (2017). A tutorial on Fisher information. *Journal of Mathematical Psychology, 80*, 40–55. doi:10.1016/j.jmp.2017.05.006

Mangat, N. S. (1994). An improved randomized response strategy. *Journal of the Royal Statistical Society. Series B (Methodological), 56*, 93–95. doi:10.1111/j.2517-6161.1994.tb01962.x

Mata, R., & Nunes, L. (2010). When less is enough: Cognitive aging, information search, and decision quality in consumer choice. *Psychology and Aging*, *25*, 289–298. doi:10.1037/a0017927

Mata, R., Schooler, L. J., & Rieskamp, J. (2007). The aging decision maker: Cognitive aging and the adaptive selection of decision strategies. *Psychology and Aging*, *22*, 796–810. doi:10.1037/0882-7974.22.4.796

Mata, R., von Helversen, B., & Rieskamp, J. (2011). When easy comes hard: The development of adaptive strategy selection. *Child Development*, *82*, 687–700. doi:10.1111/j.1467-8624.2010.01535.x

Mather, M., & Carstensen, L. L. (2005). Aging and motivated cognition: The positivity effect in attention and memory. *Trends in Cognitive Sciences*, *9*, 496–502. doi:10.1016/j.tics.2005.08.005

Matthes, T. K. (1963). On the optimality of sequential probability ratio tests. *The Annals of Mathematical Statistics*, *34*, 18–21. doi:10.1214/aoms/1177704239

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist*, *70*, 487–498. doi:10.1037/a0039400

Miller, J., & Ulrich, R. (2016). Optimizing research payoff. *Perspectives on Psychological Science*, *11*, 664–691. doi:10.1177/1745691616649170

Miller, J., & Ulrich, R. (2019). The quest for an optimal alpha. *PLOS ONE*, *14*, e0208631. doi:10.1371/journal.pone.0208631

Moors, J. J. A. (1971). Optimization of the unrelated question randomized response model. *Journal of the American Statistical Association*, *66*, 627–629. doi:10.2307/2283543

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. doi:10.1016/j.jmp.2015.11.001

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

Morey, R. D., Rouder, J. N., Verhagen, J., & Wagenmakers, E.-J. (2014). Why hypothesis tests are essential for psychological science: A comment on Cumming (2014). *Psychological Science*, *25*, 1289–1290. doi:10.1177/0956797614525969

Moshagen, M. (2010). multiTree: A computer program for the analysis of multinomial processing tree models. *Behavior Research Methods*, *42*, 42–54. doi:10.3758/BRM.42.1.42

Moshagen, M., Hilbig, B. E., Erdfelder, E., & Moritz, A. (2014). An experimental validation method for questioning techniques that assess sensitive issues. *Experimental Psychology*, *61*, 48–54. doi:10.1027/1618-3169/a000226

Moshagen, M., Hilbig, B. E., & Musch, J. (2011). Defection in the dark? A randomized-response investigation of cooperativeness in social dilemma games. *European Journal of Social Psychology*, *41*, 638–644. doi:10.1002/ejsp.793

Moshagen, M., Musch, J., & Erdfelder, E. (2012). A stochastic lie detector. *Behavior Research Methods*, *44*, 222–231. doi:10.3758/s13428-011-0144-2

Mueller, P. S., Montori, V. M., Bassler, D., Koenig, B. A., & Guyatt, G. H. (2007). Ethical issues in stopping randomized trials early because of apparent benefit. *Annals of Internal Medicine*, *146*, 878–882. doi:10.7326/0003-4819-146-12-200706190-00009

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., . . . Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*, 0021. doi:10.1038/s41562-016-0021

Neyman, J. (1941). Fiducial argument and the theory of confidence intervals. *Biometrika*, *32*, 128–150. doi:10.1093/biomet/32.2.128

Neyman, J. (1977). Frequentist probability and frequentist statistics. *Synthese*, *36*, 97–131. doi: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*, 289–337. doi:10.1098/rsta.1933.0009

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*, 615–631. doi:10.1177/1745691612459058

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, aac4716. doi:10.1126/science.aac4716

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*, 528–530. doi:10.1177/1745691612465253

Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17–59). doi:10.1016/B978-0-12-590241-0.50006-X

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

Petersen, J. L., & Hyde, J. S. (2010). A meta-analytic review of research on gender differences in sexuality, 1993–2007. *Psychological Bulletin*, *136*, 21–38. doi:10.1037/a0017504

Platt, J. R. (1964). Strong inference: Certain systematic methods of scientific thinking may produce much more rapid progress than others. *Science*, *146*, 347–353. doi:10.1126/science.146.3642.347

Pollard, P., & Richardson, J. T. (1987). On the probability of making Type I errors. *Psychological Bulletin*, *102*, 159–163. doi:10.1037/0033-2909.102.1.159

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

Psychonomic Society. (2019). Statistical Guidelines. Retrieved from https://www.psychonomic.org/page/statisticalguidelines

R Core Team. (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from http://www.R-project.org/

Read, T. R. C., & Cressie, N. A. C. (1988). *Goodness-of-fit statistics for discrete multivariate data*. New York, NY: Springer.

Reed, A. E., & Carstensen, L. L. (2012). The theory behind the age-related positivity effect. *Frontiers in Psychology*, *3*, 339. doi:10.3389/fpsyg.2012.00339

Reed, A. E., Chan, L., & Mikels, J. A. (2014). Meta-analysis of the age-related positivity effect: Age differences in preferences for positive over negative information. *Psychology and Aging*, *29*, 1–15. doi:10.1037/a0035194

Reiber, F., Schnuerch, M., & Ulrich, R. (2019). *Improving the efficiency of surveys with randomized response models: A sequential approach using curtailed sampling*. Manuscript submitted for publication.

Riefer, D. M., & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review*, *95*, 318–339. doi:10.1037/0033-295X.95.3.318

Riefer, D. M., Knapp, B. R., Batchelder, W. H., Bamber, D., & Manifold, V. (2002). Cognitive psychometrics: Assessing storage and retrieval deficits in special populations with multinomial processing tree models. *Psychological Assessment*, *14*, 184–201. doi:10.1037/1040-3590.14.2.184

Rieskamp, J., & Hoffrage, U. (2008). Inferences under time pressure: How opportunity costs affect strategy selection. *Acta Psychologica*, *127*, 258–276. doi:10.1016/j.actpsy.2007.05.004

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

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

Rouder, J. N., Morey, R. D., Verhagen, J., Province, J. M., & Wagenmakers, E.-J. (2016). Is there a free lunch in inference? *Topics in Cognitive Science*, *8*, 520–547. doi:10.1111/tops.12214

Rouder, J. N., Morey, R. D., & Wagenmakers, E.-J. (2016). The interplay between subjectivity, statistical practice, and psychological science. *Collabra*, *2*, 1–12. doi: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. doi:10.3758/PBR.16.2.225

Royall, R. (1997). *Statistical evidence: A likelihood paradigm*. Abingdon: Routledge.

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

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

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

Sanborn, A. N., & Hills, T. T. (2014). The frequentist implications of optional stopping on Bayesian hypothesis tests. *Psychonomic Bulletin & Review*, *21*, 283–300. doi: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*, 309–311. doi:10.3758/s13423-014-0607-4

Scharf, S. E., Fischer, M., & Schnuerch, M. (2020). *The attraction search effect in younger and older adults*. Manuscript in preparation.

Scharf, S. E., Wiegelmann, M., & Bröder, A. (2019). Information search in everyday decisions: The generalizability of the attraction search effect. *Judgment and Decision Making*, *14*, 488–512.

Schild, C., Heck, D. W., Ścigała, K. A., & Zettler, I. (2019). Revisiting REVISE: (Re)Testing unique and combined effects of REminding, VIsibility, and SElf-engagement manipulations on cheating behavior. *Journal of Economic Psychology*, *75*, 102161. doi:10.1016/j.joep.2019.04.001

Schnuerch, M., & Erdfelder, E. (2019). Controlling decision errors with minimal costs: The sequential probability ratio t test. *Psychological Methods.* Advance online publication. doi: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. doi:10.1016/j.jmp.2020.102326

Schnuerch, M., Heck, D. W., & Erdfelder, E. (2019). *Waldian t tests for accepting and rejecting the null hypothesis with controlled error probabilities*. Manuscript submitted for publication.

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, *25*, 128–142. doi:10.3758/s13423-017-1230-y

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*, *22*, 322–339. doi:10.1037/met0000061

Schulz, K. F., & Grimes, D. A. (2005). Sample size calculations in randomised trials: Mandatory and mystical. *The Lancet*, *365*, 1348–1353. doi:10.1016/S0140-6736(05)61034-3

Sedlmeier, P. (1996). Jenseits des Signifikanztest-Rituals: Ergänzungen und Alternativen [Beyond the ritual of significance testing: Alternative and supplementary methods]. *Methods of Psychological Research Online*, *1*, 41–63. Retrieved from https://www.dgps.de/fachgruppen/methoden/mpr-online/

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105*, 309–316. doi:10.1037/0033-2909.105.2.309

Singmann, H., & Kellen, D. (2013). MPTinR: Analysis of multinomial processing tree models in R. *Behavior Research Methods*, *45*, 560–575. doi:10.3758/s13428-012-0259-0

Smith, J. B., & Batchelder, W. H. (2008). Assessing individual differences in categorical data. *Psychonomic Bulletin & Review*, *15*, 713–731. doi:10.3758/PBR.15.4.713

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, *117*, 34–50. doi:10.1037/0096-3445.117.1.34

Söllner, A., & Bröder, A. (2016). Toolbox or adjustable spanner? A critical comparison of two metaphors for adaptive decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*, 215–237. doi:10.1037/xlm0000162

Söllner, A., Bröder, A., Glöckner, A., & Betsch, T. (2014). Single-process versus multiple-strategy models of decision making: Evidence from an information intrusion paradigm. *Acta Psychologica*, *146*, 84–96. doi:10.1016/j.actpsy.2013.12.007

Stahl, C., & Klauer, K. C. (2007). HMMTree: A computer program for latent-class hierarchical multinomial processing tree models. *Behavior Research Methods*, *39*, 267–273. doi:10.3758/BF03193157

Stallard, N., Todd, S., & Whitehead, J. (2008). Estimation following selection of the largest of two normal means. *Journal of Statistical Planning and Inference*, *138*, 1629–1638. doi:10.1016/j.jspi.2007.05.045

Tendeiro, J. N., & Kiers, H. A. L. (2019). A review of issues about null hypothesis Bayesian testing. *Psychological Methods.* Advance online publication. doi:10.1037/met0000221

Todd, P. M., & Gigerenzer, G. (Eds.). (2012). *Ecological rationality: Intelligence in the world.* Oxford ; New York: Oxford University Press.

Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin, 133*, 859–883. doi:10.1037/0033-2909.133.5.859

U.S. Census Bureau. (2019). U.S. and world population clock. Retrieved from https://www.census.gov/popclock/

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. doi:10.1027/2151-2604/a000319

Ulrich, R., Pope, H. G., Cléret, L., Petróczi, A., Nepusz, T., Schaffer, J., ... Simon, P. (2018). Doping in two elite athletics competitions assessed by randomized-response surveys. *Sports Medicine*, *48*, 211–219. doi:10.1007/s40279-017-0765-4

Ulrich, R., Schröter, H., Striegel, H., & Simon, P. (2012). Asking sensitive questions: A statistical power analysis of randomized response models. *Psychological Methods*, *17*, 623–641. doi:10.1037/a0029314

Voss, A., Rothermund, K., & Voss, J. (2004). Interpreting the parameters of the diffusion model: An empirical validation. *Memory & Cognition*, *32*, 1206–1220. doi:10.3758/BF03196893

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

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

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

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

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, *60*, 63–69. doi:10.1080/01621459.1965.10480775

Wetherill, G. B. (1975). *Sequential methods in statistics* (2. ed.). London: Chapman and Hall.

Whitehead, J. (1986). On the bias of maximum likelihood estimation following a sequential test. *Biometrika*, *73*, 573–581. doi:10.1093/biomet/73.3.573

Wrinch, D., & Jeffreys, H. (1921). On certain fundamental principles of scientific inquiry. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *42*, 369–390. doi:10.1080/14786442108633773

Yu, J.-W., Tian, G.-L., & Tang, M.-L. (2008). Two new models for survey sampling with sensitive characteristic: Design and analysis. *Metrika*, *67*, 251–263. doi:10.1007/s00184-007-0131-x

Zellner, A., & Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. *Trabajos de Estadistica Y de Investigacion Operativa*, *31*, 585–603. doi:10.1007/BF02888369

Zhang, J. J., Blumenthal, G. M., He, K., Tang, S., Cortazar, P., & Sridhara, R. (2012). Overestimation of the effect size in group sequential trials. *Clinical Cancer Research*, *18*, 4872–4876. doi:10.1158/1078-0432.CCR-11-3118

A Acknowledgements

"Wie gut", sagte der kleine Tiger, "wenn man einen Freund hat, ...
dann braucht man sich vor nichts zu fürchten." [7]

from *Oh, wie schön ist Panama*
by Janosch

This thesis would not have been possible without a number of people that guided and supported me along the way. I am grateful to each and every one of them. Let me tell you about a few who made sure I never had to be afraid.

First and foremost, I am deeply grateful to Edgar Erdfelder. He sparked in me the passion for psychological research, showed me the beauty of statistical modeling and methodological rigor, and taught me the importance of scientific writing. I thank him for bringing my attention to the SPRT, for always believing in me, and for supporting me and my ideas, no matter what. He gave me the freedom I wanted and the guidance I needed. I could not have wished for a better supervisor and mentor.

I am also grateful to Daniel Heck for his invaluable support throughout my entire PhD. His door was always open when I needed him, he always encouraged me when I was doubtful, and he always found an answer when I couldn't.

I thank Rolf Ulrich for showing me the power of a plain sheet of paper and a pencil, for sharing his ideas and wisdom on sequential analysis and the RRT, and for his invitation to Tübingen.

I am grateful to Martin Brandt, who taught me a lot about experimental psychology and computational statistics. Most importantly, he taught me R. Luckily, that was more successful than teaching me badminton. None of the projects of my dissertation would have been possible without it.

I am indebted to the research training group SMiP for giving me the opportunity to pursue my PhD in a stimulating environment and under the best possible circumstances. I thank Thorsten Meiser for inspiring and motivating all SMiPsters with his energy and spirit, for believing in me, and for encouraging me to take steps I didn't know I could. I also thank Anke Söllner and Annette Förster for their support and for making SMiP the success that it is. I thank Christoph Klauer who provided helpful advice along the way as my additional supervisor. Most importantly, I thank all my fellow

[7] *"How nice", the little tiger said, "when you have a friend, ... you don't have to be afraid of anything."*

SMiPsters for the countless wonderful after-workshop hours we spent together and for intellectual and moral support.

Pursuing a PhD can be a lonesome journey. Fortunately, I had the best travel companions one could wish for. I thank Cassandra, Elena, Feli, Leonie, Patty, and Steffen for their invaluable emotional support during this and many other journeys. I thank David, Franzi, Lili, and Nikoletta, who made sure that work always felt like home, and Michi, Philipp, and Tine, who always had an open door and a cup of coffee for me. I am especially grateful to Fabiola for making Tübingen feel like home, for helpful and inspiring conversations, and for never stopping to ask critical questions and challenging me to think; to Malte for being a great colleague, an even better musical partner, and a best friend; and to Sophie for being a light in dark times and sharing happiness when the sun is shining, as a valued colleague and an invaluable friend.

Friends are family that we choose ourselves. Luckily, my family are friends that I didn’t have to choose. I want to thank my siblings Anna, Lisa, and Robert for being my best friends in this world, my idols, and my rocks. I especially thank my brother Robert for introducing me to the wondrous world of psychology and for teaching me about science what books don’t. I thank my father for teaching me the value of hard work and rational thinking, and for his unconditional support. Finally, and most importantly, I thank my mother, to whom I owe everything. Her love, her faith in me, and her support have endured to this day and are still carrying me. For that, and for her legacy that is this family, I am forever grateful. This dissertation is for her. I like to think that it would have made her proud.

Martin Schnürch
Mannheim, January 2020

B Statement of Originality

1. I hereby declare that the presented doctoral dissertation with the title *Improving Statistical Practice in Psychological Research: Sequential Tests of Composite Hypotheses* is my own work.
2. I did not seek unauthorized assistance of a third party and I have employed no other sources or means except the ones listed. I clearly marked any quotations derived from the works of others.
3. I did not present this doctoral dissertation or parts of it at any other higher education institution in Germany or abroad.
4. I hereby conform the accuracy of the declaration above.
5. I am aware of the significance of this declaration and the legal consequences in case of untrue or incomplete statements.

I affirm in lieu of oath that the statements above are to the best of my knowledge true and complete.

Signature:

Date:

C Co-Authors' Statements

Co-Author: Edgar Erdfelder

I hereby confirm that the following articles included in the thesis *Improving Statistical Practice in Psychology: Sequential Tests of Composite Hypotheses* were primarily conceived and written by Martin Schnürch, PhD candidate in the DFG Research Training Group "Statistical Modeling in Psychology" at the University of Mannheim:

Schnuerch, M., & Erdfelder, E. (2019). Controlling decision errors with minimal costs: The sequential probability ratio *t* test. *Psychological Methods*. Advance online publication. http://dx.doi.org/10.1037/met0000234

Schnuerch, M., Heck, D. W., & Erdfelder, E. (2019). *Waldian t tests for accepting and rejecting the null hypothesis with controlled error probabilities.* Manuscript submitted for publication.

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

I sign this statement to the effect that Martin Schnürch is credited as the primary source of the ideas and the main author of all three articles. He devised the theoretical and statistical background of the sequential procedures, implemented them in statistical software, performed the simulations and analyses, wrote the first drafts, and contributed to improving and revising the manuscripts. I contributed to developing and refining the theoretical background, provided suggestions for the simulated scenarios, advised the development and power analyses of the MPT-model examples in the last article, and provided recommendations for structuring and improving the manuscripts.

Prof. Dr. Edgar Erdfelder
Mannheim, January 2020

Co-Author: Daniel W. Heck

I hereby confirm that the following articles included in the thesis *Improving Statistical Practice in Psychology: Sequential Tests of Composite Hypotheses* were primarily conceived and written by Martin Schnürch, PhD candidate in the DFG Research Training Group "Statistical Modeling in Psychology" at the University of Mannheim:

Schnuerch, M., Heck, D. W., & Erdfelder, E. (2019). *Waldian t tests for accepting and rejecting the null hypothesis with controlled error probabilities.* Manuscript submitted for publication.

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

I sign this statement to the effect that Martin Schnürch is credited as the primary source of the ideas and the main author of both articles. He worked out the theoretical and statistical background, implemented the developed sequential procedures in statistical software, performed the simulations and analyses, wrote the first drafts, and contributed to improving and revising the manuscripts. I contributed to refining the mathematical derivation of the procedures and clarifying their presentation in the manuscripts, provided recommendations for the simulation settings, suggested the hypothetical experimental context for the examples in the last article, and contributed to improving the manuscripts.

Prof. Dr. Daniel W. Heck
Marburg, January 2020

D Copies of Articles

Controlling Decision Errors with Minimal Costs: The Sequential Probability Ratio *t* Test

Martin Schnuerch and Edgar Erdfelder
University of Mannheim

The official citation that should be used in referencing this material is: Schnuerch, M., & Erdfelder, E. (2019). Controlling decision errors with minimal costs: The sequential probability ratio t test. *Psychological Methods*. http://dx.doi.org/10.1037/met0000234.

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

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

Critical tests of theories and hypotheses are at the heart of psychological science. A good theory makes clear-cut predictions that can be evaluated empirically, for example, in an experiment. Empirical tests of such predictions often take the form of binary decisions: Based on the data, do we accept the hypothesis of interest or do we reject it, thereby corroborating or refuting the underlying theory? The most common statistical procedure in psychology to decide between conflicting hypotheses is usually referred to as *null-hypothesis significance testing* (NHST). NHST has been harshly criticized in the past, and rightly so, as it is an inconsistent hybrid between two seemingly similar but in fact substantially different statistical theories: the theory of null-hypothesis testing proposed by Fisher, and the theory of statistical decision-making by Neyman and Pearson (e.g., Bakan, 1966; Berger, 2003; Bredenkamp, 1972; Cumming, 2014; Dienes, 2011; Gelman, 2016; Gigerenzer, 1993, 2004; Goodman, 1993; Sedlmeier, 1996; Wagenmakers, 2007). Notwithstanding these criticisms, NHST has been the dominant procedure in behavioral science for decades. However, fostered by the reproducibility crisis in psychology (Asendorpf et al., 2013; Earp & Trafimow, 2015; Maxwell, Lau, & Howard, 2015; Open Science Collaboration, 2015; Pashler & Wagenmakers, 2012; but see Gilbert, King, Pettigrew, & Wilson, 2016), there is an increasing awareness of the pitfalls of NHST and the importance of rigorous control of decision errors in hypothesis testing.

According to Neyman and Pearson (1933), two types of errors can occur when deciding between a null hypothesis ($\mathcal{H}_0$) and an alternative hypothesis ($\mathcal{H}_1$): The null hypothesis is rejected when it is true (Type 1 error), or it is accepted when it is false (Type 2 error). By convention, the probabilities of Type 1 and 2 errors are denoted by α and β, respectively. The complement of β, $1-\beta$, is referred to as the statistical power of the test. As outlined in the statistical guidelines of the Psychonomic Society, "[i]t is important to address the issue of statistical power. [...] Studies with low statistical power produce inherently ambiguous results because they often fail to replicate" (Psychonomic Society, 2012). Despite such pleas, however, the issue of power has largely been neglected in psychological research so far. A possible reason is that the most common statistical procedure to control α and β (i.e., the Neyman-Pearson procedure) often requires sample sizes much larger than those typically employed (Erdfelder, Faul, & Buchner, 1996). To illustrate, a two-tailed two-sample t test requires a total sample size of $N = 210$ to detect a mean difference of medium size (i.e., Cohen's $d = .50$) with error probabilities $\alpha = \beta = .05$. In contrast, the common overall sample size for the same test is only about $N = 60$ in prototypical journal publications, resulting in power t test slightly lower than $1 - \beta = .50$ (Cohen, 1962; Sedlmeier & Gigerenzer, 1989).

To avoid costly hypothesis tests, researchers may be tempted to apply NHST to small, underpowered samples first, followed by recursive increases in sample size until a significant test result is observed. This misleading use of NHST is known as data peeking, a questionable research practice that boosts chances of gaining a significant outcome at the cost of error probability control (Simmons, Nelson, & Simonsohn, 2011). In this article, we promote a proper alternative statistical method that was developed more than 70 years ago: Sequential Analysis (Wald, 1947). Unlike data peeking with its associated risk of inflating Type 1 errors, sequential hypothesis tests have been designed specifically to control error probabilities while at the same time allowing for smaller sample sizes than the Neyman-Pearson approach (Lakens, 2014). As computational tools have improved substantially over the past decades, these sequential tests are nowadays easily implemented and combined with standard statistical software. We will empirically demonstrate the beneficial properties of one particular sequential test, namely, the Sequential Probability Ratio Test (Wald, 1947). Moreover, we will show that on top of controlling for decision error probabilities, this test is more efficient than both the Neyman-Pearson approach and other common sequential designs. Importantly, we will also assess the robustness of the proposed sequential test against violations of its assumptions.

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

One helpful step in this direction was recently taken by Schönbrodt, Wagenmakers, Zehetleitner, and Perugini (2017). These authors proposed a sequential method based on Bayesian inference, referred to as Sequential Bayes Factors (SBFs). By means of simulation, they demonstrated the properties of SBFs in the context of testing hypotheses about mean differences of two

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

This research was supported by a grant from the Deutsche Forschungsgemeinschaft (DFG, GRK 2277) to the Research Training Group "Statistical Modeling in Psychology" (SMiP).

Parts of this manuscript were presented at the 60th Conference for Experimental Psychologists (2018) in Marburg, Germany.

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

Correspondence concerning this article should be addressed to Martin Schnuerch or Edgar Erdfelder, Cognition and Individual Differences Lab, Department of Psychology, University of Mannheim, 68131 Mannheim, Germany. E-mail: martin.schnuerch@psychologie.uni-mannheim.de

independent groups (two-sample *t* test). Specifically, they simulated populations with a specific mean difference δ and examined the simulation estimate of the expected sample size and the relative frequencies of Type 1 and Type 2 errors of SBFs for different prior specifications and stopping criteria. Based on their simulations, they compared the SBF design to two other designs: the standard fixed-sample Neyman-Pearson *t* test (misleadingly referred to as null-hypothesis significance test with power analysis [NHST-PA]) and the group sequential (GS) design (Proschan, Lan, & Wittes, 2006).

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

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

In the following section, we briefly outline the basic concept of the SPRT with particular focus on its application to the *t*-test scenario, elaborating on differences between the SPRT and other sequential designs. Next, we evaluate by means of simulation the properties of the SPRT with regard to empirical rates of incorrect decisions. In Section 4, we empirically compare SPRT, GS, and SBFs in terms of efficiency, that is, the expected sample size required to reach a decision. Subsequently, we explore the effects of violations of various assumptions underlying the test procedures. We then demonstrate the SPRT using an empirical example, and discuss implications as well as limitations of our study and the SPRT. Finally, we provide recommendations on how to apply the proposed sequential *t* test in research practice.

The Sequential Probability Ratio Test

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

$$\begin{aligned}&\text{1) Accept } \mathcal{H}_1 \text{ and reject } \mathcal{H}_0;\\ &\text{2) Accept } \mathcal{H}_0 \text{ and reject } \mathcal{H}_1;\\ &\text{3) Continue sampling.}\end{aligned} \tag{1}$$

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

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

[1] Average Sample Number denotes the average number of observations per sample, that is, the expected sample size at termination. Wald (1947) consistently used this expression, thus, we will maintain it as a technical term throughout the article.

(Wald, 1947):

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

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

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

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

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

Consistency: The likelihood ratio is consistent, that is, if one of the specified hypotheses is in fact true, it will converge to either 0 or ∞ as the sample size increases towards infinity. Note that not all tests actually behave in this reasonable way. The p value in an NHST, for example, will not converge to 1 if the null hypothesis is true, which is why it is not suitable as a measure of evidence for the null (Rouder, Speckman, Sun, Morey, & Iverson, 2009).

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

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

1) Accept $\mathcal{H}_1$ and reject $\mathcal{H}_0$ when $LR_m \geq A$;
2) Accept $\mathcal{H}_0$ and reject $\mathcal{H}_1$ when $LR_m \leq B$; (4)
3) Sample a new independent observation x_{m+1} when $B < LR_m < A$.

Wald (1947) showed that this sequential procedure terminates with probability 1 after a finite number of observations with either decision 1) or 2). This implies that $A \leq P(\text{accept } \mathcal{H}_1|\theta_1)/P(\text{accept } \mathcal{H}_1|\theta_0)$ and $B \geq P(\text{accept } \mathcal{H}_0|\theta_1)/P(\text{accept } \mathcal{H}_0|\theta_0)$ (Wetherill, 1975). For practical purposes, these inequalities can be replaced by equalities and, in accordance with the requirements given in (2), the boundaries may simply be determined by $A = (1-\beta)/\alpha$ and $B = \beta/(1-\alpha)$. The resulting test will be approximately of strength (α, β): As the test statistic may exceed one of the boundaries at the point of termination rather than matching it exactly (a phenomenon called "overshooting"), the actual error

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

probabilities of the sequential procedure will in general be lower than α and β. Hence, strictly speaking, the SPRT is an approximate test with α and β serving as upper bounds to the error probabilities.

Importantly, this also holds true for interval hypotheses of the form $\mathcal{H}_0$: $\theta \leq \theta_0$ versus $\mathcal{H}_1$: $\theta \geq \theta_1$ ($\theta_0 < \theta_1$) if all other parameters of the statistical model are known constants. Like the classical Neyman-Pearson test, an SPRT based on the simple hypotheses $\theta = \theta_0$ versus $\theta = \theta_1$ will have its maximum error probabilities α and β if the true θ equals θ_0 and θ_1, respectively. For any other true value θ in line with $\mathcal{H}_0$: $\theta \leq \theta_0$ or $\mathcal{H}_1$: $\theta \geq \theta_1$ ($\theta_0 < \theta_1$), the respective error probabilities will be lower (Wald, 1947). Hence, just like Neyman-Pearson tests, SPRTs allow for the specification of upper-bound error probabilities even if there is no expectation of the exact value of the parameter of interest, as long as a minimum (maximum) value can be defined and all other parameters are constants.

Sequential *t* Tests

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

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

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

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

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

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

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

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

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

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

$$LR_m = \frac{f(t^2_m | df_m, \Delta_m)}{f(t^2_m | df_m)}, \qquad (8)$$

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

$$\Delta_m = d \cdot \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}}. \qquad (9)$$

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

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

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

Both the t and the F density function are available in the standard R environment (R Core Team, 2017). Thus, an SPRT for a one- or two-sample t test can be conducted easily with R by iteratively computing the ratios given in (5) for a one-sided test, and (10) for a two-sided test for each stage m of the sequential sampling process. A workable R script to apply the SPRTs described in this section can be downloaded from the Open Science Framework (https://osf.io/4zub2/).

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

Therefore, it is of great practical as well as theoretical interest to empirically assess the properties of Hajnal's t test and examine (1) the degree to which the actual error rates approximate the upper bounds α and β, (2) the expected sample size and relative efficiency as compared with Schönbrodt et al.'s SBFs and the GS design, and (3) the robustness of these results when basic assumptions are violated. In the following section, we will elaborate on the differences between the SPRT and the two alternative sequential test procedures addressed in this article.

Two Alternative Sequential Designs: GS and SBFs

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

Due to the interim analyses and the resulting possibility to terminate early, the GS requires on average fewer observations than Neyman-Pearson tests with the same error probabilities (Schönbrodt et al., 2017). Importantly, it allows for explicit control of these probabilities and the specification of a maximum number of observations required. As the interim analyses have to be planned a priori, however, the GS is less flexible than the SPRT or SBFs. Whereas the latter allow for termination after possibly any single additional observation, a GS test can only terminate at one of the planned stops. Hence, although it has the advantage of a definite upper limit to the required sample size, it can be expected that the GS is on average less efficient than SPRT and SBFs (see Schönbrodt et al., 2017).

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

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

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

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

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

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

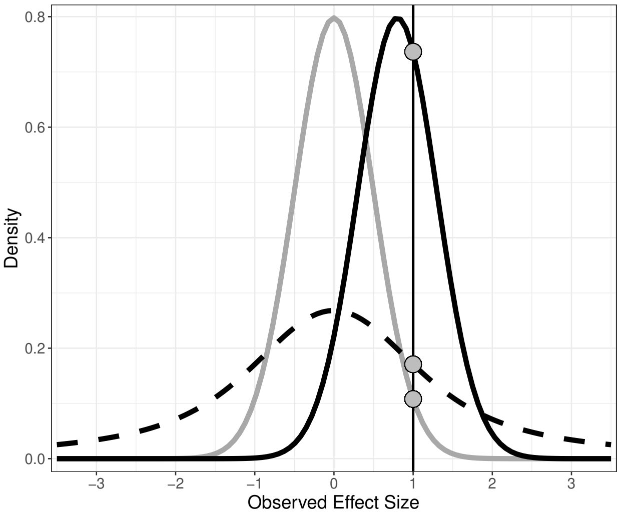

Figure 1. Effects of marginalizing across an effect-size prior, assuming some known variance. The grey line denotes the probability density of an observed effect size under the null hypothesis $\delta = 0$. The solid black line denotes the probability density of observed data under the alternative hypothesis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

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

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

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

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

It should be noted, however, that this sensitivity comes at a cost: If the true effect differs greatly from what was expected ($\hat{\delta} = 3$, say), the likelihood ratio for the point alternative hypothesis will be less pronounced than for the diffuse hypothesis. As a consequence, in such a case the SPRT is likely to be less efficient, while an SBF based on a diffuse prior will be more robust. A similar point was recently made by Stefan, Gronau, Schönbrodt, and Wagenmakers (2019). According to these authors, an SBF based on an informative prior is more efficient (or less error-prone) when the true effect lies within the prior's highest density region. At the same time, however, the informative prior might be at a disadvantage if the true effect greatly deviates from this region. In other words, there is a general trade-off between peak efficiency when the true effect matches the expectation, and robustness when it doesn't. Conceptually, the effect size specified in the SPRT is the most extreme case of an informed prior. Hence, Stefan et al.'s conclusions also apply to the SPRT.

Statistical Error Rates of SPRT, GS, and SBFs

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

Note, however, that whereas Hajnal's t test and the GS design are based on the assumption of a fixed underlying effect and the same nominal error rates, the default priors in the SBFs make quite different assumptions. In a Cauchy distribution with scale parameter r, 50% of the area under the curve lie in the interval $[-r, r]$. Thus, the default scale parameters used by Schönbrodt et al. (2017), $r = \sqrt{2}/2$, 1, and $\sqrt{2}$, correspond to expected median absolute effect sizes of $\delta = 0.7$, 1, and 1.4, respectively. The absolute effect sizes corresponding to a small, medium, or large effect in Hajnal's t test as well as GS and Neyman-Pearson tests, in contrast, are $\delta = 0.2$, 0.5, and 0.8, respectively (Cohen, 1988). Thus, our results—like those of Schönbrodt et al. (2017)—should not be generalized to other SBF designs with different prior distributions (e.g., informative priors; Stefan et al., 2019) or other population scenarios (e.g., random effects).

Settings of the Simulation

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

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

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

Results

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

For the sake of brevity, we only display a limited range of effect sizes here, namely, $\delta = 0$, 0.2, 0.5, and 0.8, as these represent the absence of an effect and the effect sizes commonly referred to as small, medium, and large (Cohen, 1988). In a similar vein, we only report a subset of SBF threshold *t* test, namely 5, 10, and 30. The full table of results as well as reproducible scripts and all data can be downloaded from https://osf.io/4zub2/. The ASN (as well as the 50th, 75th, and 95th quantile) for Hajnal's *t* test, GS, and SBFs corresponding to the results displayed in Table 1 may be obtained from the Appendix (Table A1).

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

We observe a similar result for the GS: The empirical error rates nicely approximate the nominal α. In some cases, the estimate is slightly above the nominal level, but this is likely caused by sampling error. Hence, with respect to Type-1 error control, the GS and the SPRT procedures are comparable and perform well.

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

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

Notably, this result also holds when the true effect does not match the expected effect but is in fact larger.

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

Table 1

Percentages [and 95% CI] of Type 1 and Type 2 Decision Errors Committed by Hajnal's t Test, Group Sequential Test, and Sequential Bayes Factors

		Hajnal's *t* test				Group Sequential Test				Sequential Bayes Factors		
		$\alpha = 1\%$		$\alpha = 5\%$		$\alpha = 1\%$		$\alpha = 5\%$				
d	r	$\beta = 5\%$	$\beta = 10\%$	$\beta = 5\%$	$\beta = 10\%$	$\beta = 5\%$	$\beta = 10\%$	$\beta = 5\%$	$\beta = 10\%$	$BF = 5$	$BF = 10$	$BF = 30$
						$\delta = 0$ (% Type 1 error)						
0.2	$\sqrt{2}/2$	1.0 [0.9,1.3]	0.9 [0.7,1.1]	4.5 [4.1,4.9]	4.8 [4.4,5.3]	0.8 [0.7,1]	1.0 [0.8,1.2]	5.1 [4.6,5.5]	4.7 [4.3,5.1]	11.6 [11.0,12.3]	6.3 [5.9,6.8]	2.3 [2.0,2.6]
0.5	1	0.8 [0.7,1.0]	0.7 [0.5,0.9]	4.5 [4.1,5.0]	3.8 [3.4,4.2]	0.8 [0.7,1.0]	0.9 [0.7,1.1]	5.3 [4.9,5.8]	4.9 [4.5,5.3]	10.3 [9.7,10.9]	5.8 [5.3,6.3]	2.0 [1.8,2.3]
0.8	$\sqrt{2}$	0.8 [0.7,1.0]	0.8 [0.6,1.0]	4.0 [3.6,4.4]	3.8 [3.5,4.2]	1.2 [1.0,1.5]	1.3 [1.1,1.5]	5.4 [5.0,5.9]	5.6 [5.1,6.0]	10.5 [9.9,11.1]	5.8 [5.4,6.3]	2.1 [1.8,2.4]
						$\delta = 0.2$ (% Type 2 error)						
0.2	$\sqrt{2}/2$	4.7 [4.3,5.1]	9.4 [8.8,10.0]	4.4 [4.0,4.8]	9.1 [8.5,9.6]	4.4 [4.0,4.8]	9.7 [9.1,10.3]	5.1 [4.7,5.6]	9.1 [8.6,9.7]	49.0 [48.0,50.0]	6.4 [5.9,6.9]	0.0 [0.0,0.0]
0.5	1	81.0 [80.2,81.7]	84.8 [84.1,85.5]	73.1 [72.2,73.9]	74.6 [73.7,75.5]	80.5 [79.7,81.3]	84.2 [83.4,84.9]	68.2 [67.3,69.2]	73.6 [72.8,74.5]	67.0 [66.0,67.9]	23.5 [22.7,24.4]	0.0 [0.0,0.0]
0.8	$\sqrt{2}$	95.2 [94.7,95.6]	95.3 [94.9,95.7]	88.1 [87.5,88.8]	89.2 [88.5,89.8]	92.7 [92.1,93.2]	93.8 [93.4,94.3]	85.2 [84.5,85.9]	85.6 [84.9,86.3]	78.6 [77.8,79.4]	49.0 [48.1,50.0]	0.1 [0.1,0.2]
						$\delta = 0.5$ (% Type 2 error)						
0.2	$\sqrt{2}/2$	0.0 [0.0,0.0]	0.0 [0.0,0.0]	0.0 [0.0,0.0]	0.0 [0.0,0.0]	0.0 [0.0,0.0]	0.0 [0.0,0.0]	0.0 [0.0,0.0]	0.0 [0.0,0.0]	1.7 [1.4,1.9]	0.0 [0.0,0.0]	0.0 [0.0,0.0]
0.5	1	4.2 [3.8,4.6]	8.8 [8.2,9.4]	4.5 [4.1,4.9]	8.7 [8.2,9.3]	4.2 [3.8,4.6]	8.6 [8.1,9.2]	4.7 [4.2,5.1]	8.5 [8.0,9.1]	11.7 [11.1,12.3]	0.0 [0.0,0.1]	0.0 [0.0,0.0]
0.8	$\sqrt{2}$	39.8 [38.8,40.8]	48.8 [47.8,49.8]	35.4 [34.4,36.3]	43.6 [42.6,44.6]	43.5 [42.5,44.5]	52.9 [51.9,53.9]	35.4 [34.5,36.4]	44.7 [43.7,45.6]	31.8 [30.9,32.7]	1.3 [1.1,1.5]	0.0 [0.0,0.0]
						$\delta = 0.8$ (% Type 2 error)						
0.2	$\sqrt{2}/2$	0.0 [0.0,0.0]	0.0 [0.0,0.0]	0.0 [0.0,0.0]	0.0 [0.0,0.0]	0.0 [0.0,0.0]	0.0 [0.0,0.0]	0.0 [0.0,0.0]	0.0 [0.0,0.0]	0.0 [0.0,0.0]	0.0 [0.0,0.0]	0.0 [0.0,0.0]
0.5	1	0.0 [0.0,0.1]	0.1 [0.0,0.2]	0.0 [0.0,0.1]	0.2 [0.1,0.3]	0.0 [0.0,0.0]	0.0 [0.0,0.1]	0.0 [0.0,0.1]	0.1 [0.0,0.2]	0.6 [0.4,0.8]	0.0 [0.0,0.0]	0.0 [0.0,0.0]
0.8	$\sqrt{2}$	3.9 [3.5,4.3]	7.7 [7.2,8.2]	4.1 [3.7,4.5]	8.2 [7.7,8.8]	4.5 [4.1,4.9]	8.0 [7.4,8.5]	4.1 [3.7,4.5]	8.7 [8.2,9.3]	5.9 [5.4,6.4]	0.0 [0.0,0.1]	0.0 [0.0,0.0]

Note. The d and r metrics indicate fundamentally different effect-size expectations, although they are often assigned the same verbal labels for "small" ($d = 0.2$, $r = \sqrt{2}/2$), "medium" ($d = 0.5$, $r = 1$), and "large" effects ($d = 0.8$, $r = \sqrt{2}$). The group sequential test comprised three interim and one final test. d = expected effect size according to $\mathcal{H}_1$ in Hajnal's t test and group sequential test (Cohen's d); r = scale parameter of Cauchy prior (= expected median absolute effect size) according to $\mathcal{H}_1$ in Sequential Bayes Factors; δ = true population effect size (Cohen's d in population); BF = threshold Bayes factor.

As Table 1 shows, the probability that Hajnal's t test incorrectly accepts a false null hypothesis converges to 0 when $\delta > d$. It is a popular critique by proponents of the Bayesian approach that a precise prediction of the effect size is not possible (e.g., Schönbrodt et al., 2017). Even in this case, however, a test can be defined with β as an upper bound to the Type 2 error probability (Wald, 1947). By specifying a minimum relevant effect d_{min} and setting up Hajnal's t test for the simple hypothesis $\mathcal{H}_0$: $\delta = 0$ against $\mathcal{H}_1$: $\delta = d_{min}$, the probability of incorrectly accepting a false null hypothesis will never exceed β if $\delta \geq d_{min}$. Of course, if the true effect is notably smaller than d_{min} then the probability of accepting $\mathcal{H}_0$ will exceed β. However, if d_{min} is specified based on which effect sizes are practically relevant, one would actually prefer the test to maintain $\mathcal{H}_0$ if the true effect falls under this lower bound. Thus, the results demonstrate that Hajnal's t test provides reliable, conservative control of the probabilities to commit a decision error.

It should be noted at this point, however, that a conservative specification of the effect does not only result in conservative error rates but also in a less efficient test: In the same way as error rates decrease when the true effect is larger than expected, the ASN increases (Table A1). This is not surprising as this reflects the trade-off between efficiency and robustness (Stefan et al., 2019). Importantly, this is also true for the GS and the classical Neyman-Pearson test. As the effect-size assumption is the same for all three designs, a conservative estimate will increase the required sample size for all of them. As Table A1 shows, however, Hajnal's t test is still more efficient in these cases. For example, if a small effect is expected ($d = 0.2$) in case of a medium true effect ($\delta = 0.5$), Hajnal's t test with $\alpha = \beta = .05$ requires on average 194 observations. The GS test with the same parameters requires on average 378 observations. A classical t test with the same assumptions would even require 1302 observations. Hence, Hajnal's t test is not only more efficient when the correct effect size is expected, but also when the tests specify a conservative assumption.

As in case of Type 1 errors, there is no explicit control of Type 2 errors in the SBF design. For a threshold Bayes factor of 5, the empirical error rates exceed typical error rates by far (see also Schönbrodt et al., 2017). A more reasonable threshold of 10 yields excellent error probabilities for medium to large effects but not for small effects ($\delta = 0.2$). If a higher threshold is chosen ($BF = 30$), the procedure will basically commit no decision errors, even when the true effect is small. However, this powerful procedure comes at the cost of efficiency: In the context of small to medium effect sizes, the expected sample sizes required to reach the decision threshold can become extremely large (see Table A1). For example, for a true effect of size $\delta = 0.2$, an SBF assuming a Cauchy prior with scale $r = \sqrt{2}/2$ requires on average 1120 observations to reach a threshold of 30. To summarize, the results indicate that the SBF design—if combined with thresholds representing moderate ($BF = 5$) or strong evidence ($BF = 10$)—can be associated with high error probabilities and lacks a proper means to control these explicitly.

Relative Efficiency of SPRT, GS, and SBFs

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

Settings of the Simulations

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

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

Results

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

[5] In conditions for which the SBFs did not exhibit wrong decisions, β was set to an arbitrarily small value of 1/50000.

smaller than N_{NP}. In almost all conditions, the observed ASN undercuts the corresponding statistics of the SBFs and the GS.

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

(1) As the true effect further departs from the null, the likelihood of an observation that is typical under the null hypothesis decreases quickly. Thus, the expected change in the likelihood ratio by adding a single observation increases correspondingly fast. Therefore, both sequential procedures reach a decision on average after very few observations already (e.g., for $\delta = 1.2$, all ASN are smaller than 20). Hence, the comparison is distorted in this context by a floor effect.

(2) The extent to which the likelihood ratio exceeds the stopping t test A and B at the point of termination (*overshooting*) increases as a function of effect size (Wald, 1947). Thus, with increasing effect size, the actual error probabilities of the SPRT further depart from the specified α and β, resulting in a more conservative and slightly less efficient test. Wald conjectured that this loss of efficiency was not of practical relevance. Nevertheless, it is important to keep in mind when interpreting the small difference in efficiency between Hajnal's t test and SBFs in the context of large effect sizes.

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

Robustness of SPRT, GS, and SBFs

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

H. Lee and Fung (1980) already examined the robustness of Hajnal's t test under conditions of non-normality and heteroscedasticity. Due to computational limitations at the time, however, their simulations were based on approximations to the likelihood ratio. In this section, we examine the performance of Hajnal's t test as well as the GS and the default SBFs under conditions of (1) non-normality and (2) heteroscedasticity, as well as (3) random effects and (4) intentional misuse. For the sake of parsimony, we restricted the simulations to the nominal error rates $\alpha = \beta = .05$ in Hajnal's t test and the GS. For the SBFs, we chose a threshold value of $BF = 10$ throughout the simulations. As this value reflects "strong evidence" from a Bayesian perspective (M. D. Lee & Wagenmakers, 2013), it is often used as a threshold in practical applications (e.g., Matzke et al., 2015; Schönbrodt et al., 2017; Wagenmakers et al., 2015). In each simulation, 10,000 replications per parameter combination were simulated. All scripts and data are again available from the Open Science Framework.

Non-Normality

Settings. To investigate the test procedures' performances against violations of the normality assumption we repeated the first simulation for data generated from log-normal distributions and mixtures of two normal distributions. For the former case, we drew random data for two groups from a log-normal distribution corresponding to a standard normal on the log scale. To each observation in the first group, $\delta\sigma'$ was added, where δ denotes the true standardized mean difference ($\delta = 0$, 0.2, 0.5, 0.8) and σ' represents the standard deviation of the log-normal distribution.

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

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

where γ ($\gamma = .9, .7, .5$) denotes the probability that an observation is drawn from $\mathcal{N}(\mu_1, \sigma_1)$. For the underly-

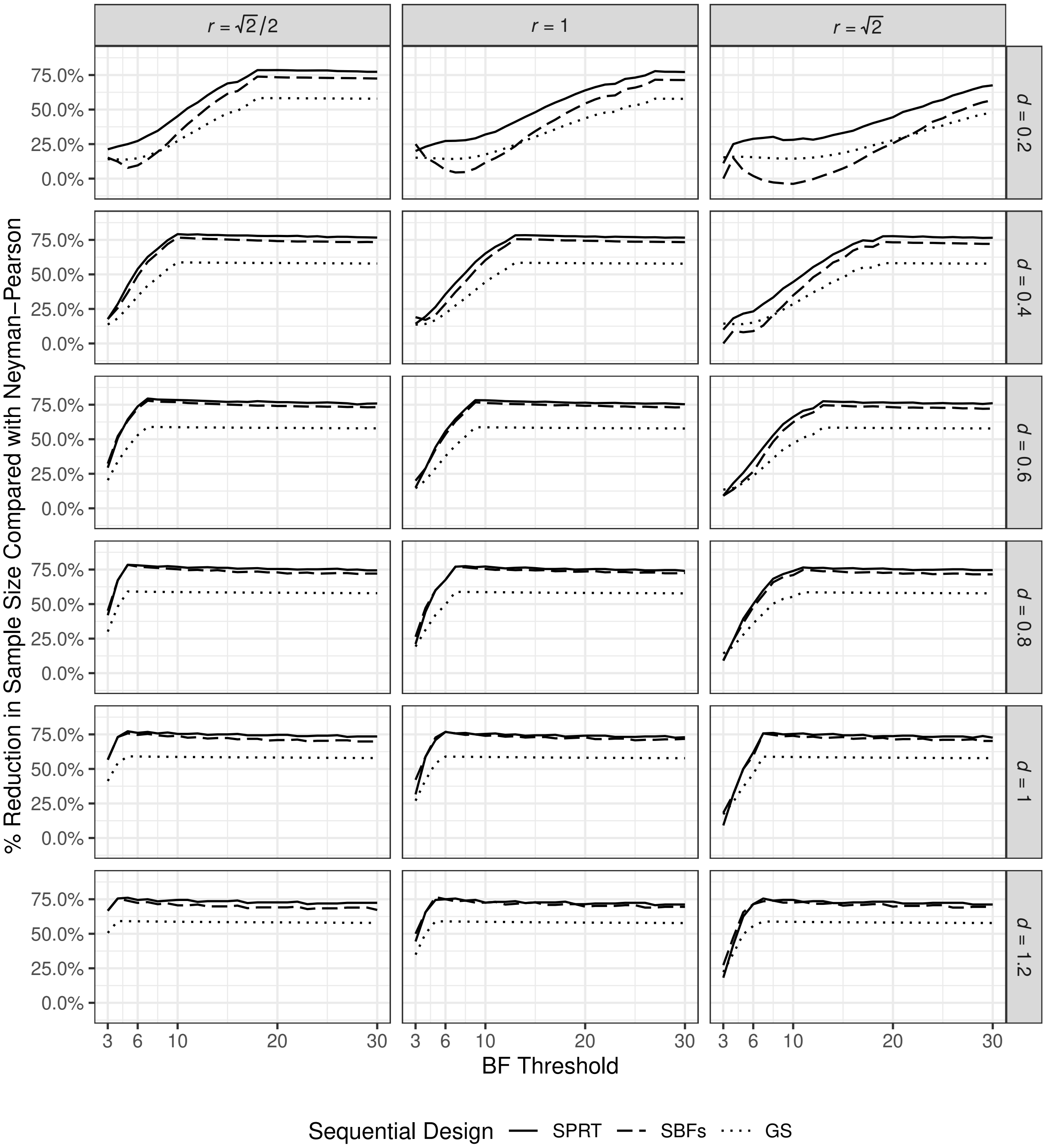

Figure 2. Relative efficiency of sequential probability ratio test (SPRT), group sequential test (GS) and sequential Bayes factors (SBFs). The y-axis denotes the reduction in expected sample size of SPRT (solid line), GS (dotted line) and SBFs (dashed line) compared with a Neyman-Pearson *t* test with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

Table 2

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

					$\delta = 0.2$			$\delta = 0.5$			$\delta = 0.8$		
Distribution	γ	s	k	Empirical error rates	SPRT	GS	SBF	SPRT	GS	SBF	SPRT	GS	SBF
Normal		0.0	3.0	α'	4.5	5.1	6.3	4.5	5.3	5.8	4.0	5.4	5.8
				β'	4.4	5.1	6.4	4.5	4.7	0.0	4.1	4.1	0.0
Mixture	.9	0.8	6.0	α'	4.5	3.3	5.5	3.7	4.3	5.1	3.5	4.4	5.1
				β'	4.5	3.9	5.6	3.9	3.7	0.0	4.1	4.2	0.0
	.7	0.9	4.4	α'	4.6	2.5	6.2	4.2	2.5	4.6	3.7	3.1	4.5
				β'	2.9	2.0	4.2	2.9	1.8	0.0	2.6	1.8	0.0
	.5	0.7	3.4	α'	2.3	1.2	5.2	3.8	1.5	5.1	4.0	1.7	4.7
				β'	0.8	0.4	1.6	1.3	0.4	0.0	1.4	0.3	0.0
Log-normal		6.2	116.9	α'	3.9	4.8	4.0	2.7	4.7	3.7	1.8	4.6	3.3
				β'	4.6	5.1	6.2	4.1	4.5	0.1	3.7	4.5	0.0

Note. The first two rows display results from the first simulation for normally distributed data, see Table 1, columns 3, 7, and 10. Number of repetitions per parameter combination: $k = 10,000$. γ = mixture probability; s = skewness; k = kurtosis; δ = true and expected effect size (Cohen's d in population); $SPRT$ = sequential probability ratio test (Hajnal's t test) assuming $d = \delta$ and $\alpha = \beta = .05$. GS = group sequential design with four tests, assuming $d = \delta$ and $\alpha = \beta = .05$. SBF = sequential Bayes factor design with threshold 10, assuming r = $\sqrt{2}/2$, 1, $\sqrt{2}$ when δ = 0.2, 0.5, 0.8, respectively.

ing distributions, we defined $\mu_1 = 0$, $\sigma_1 = 1$, $\mu_2 = 2$, and $\sigma_2 = 2$. As in the log-normal case, $\delta\sigma'$ was added to each observation in the first group, with σ' denoting the standard deviation of the mixture distribution.

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

In terms of error rates, the examined procedures are quite robust against violations of distributional assumptions. This is not surprising as it is in line with H. Lee and Fung's results for Hajnal's t test. Type 1 error rates in particular seem to be quite stable for all designs across all simulated scenarios, although Hajnal's t test becomes slightly conservative with increasing effect-size assumption for log-normally distributed data. In the case of mixture distributions, there is a tendency that Type 2 error rates for all test procedures decrease with decreasing kurtosis. Interestingly, however, this is not accompanied by an increase in expected sample sizes. To summarize, all three sequential designs show robustness under conditions of non-normality both in terms of error rates and ASN.

Heteroscedasticity

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

$$\mu_2 = \delta \cdot \sqrt{\frac{\sigma_2^2 + 1}{2}} \qquad (14)$$

with $\sigma_2 = 1/4, 1, 4$ and $\delta = 0, 0.2, 0.5, 0.8$. In addition, we simulated two sampling schemes (H. Lee & Fung, 1980): (1) pairwise sampling from the two populations such that $n_1/n_2 = 1$, and (2) unbalanced sampling such that at each step for one observation in the first sample there were always three in the second sample, that is, $n_1/n_2 = 1/3$.

Results. The observed error rates for the three test procedures under the condition of heteroscedasticity are displayed in Table 3. The corresponding expected sample sizes can be obtained from the Appendix (Table A3). If the sample sizes are balanced, Hajnal's t test is basically unaffected by heteroscedasticity in the underlying populations. Although there seems to be a slight tendency that with increasing expected effect size, the Type 1 error rate increases as well, this is likely due to sampling error. In the same vein, expected sample sizes of Hajnal's t test are basically constant irrespective of

Table 3

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

			$\delta = 0.2$			$\delta = 0.5$			$\delta = 0.8$		
N_1/N_2	σ_1/σ_2	Empirical error rates	SPRT	GS	SBF	SPRT	GS	SBF	SPRT	GS	SBF
1	1/4	α'	4.6	4.9	9.8	5.1	5.0	9.5	5.3	6.3	8.6
		β'	4.6	4.8	6.2	4.3	4.5	0.0	3.5	4.0	0.0
	1	α'	4.3	5.2	6.5	4.0	5.3	6.0	4.1	5.6	5.5
		β'	4.8	4.8	6.0	4.4	4.7	0.0	4.1	4.0	0.0
	4	α'	4.8	5.7	9.8	5.0	5.4	9.3	5.3	6.5	9.3
		β'	4.6	4.5	5.9	4.6	4.5	0.0	3.7	4.0	0.0
1/3	1/4	α'	0.0	1.0	0.1	0.0	0.8	0.2	0.1	1.1	0.1
		β'	1.2	17.2	1.9	1.0	17.1	0.0	1.1	16.3	0.0
	1	α'	4.4	5.1	6.0	3.8	5.0	5.4	3.1	5.6	4.3
		β'	4.6	14.5	6.1	4.0	14.4	0.0	3.4	13.1	0.0
	4	α'	38.0	21.3	59.8	36.1	23.7	55.9	34.9	26.6	52.3
		β'	6.1	10.2	5.0	5.4	9.8	0.1	4.9	9.0	0.0

Note. Number of repetitions per parameter combination: $k = 10,000$. N_1/N_2 = ratio of sample sizes in group 1 and 2; σ_1/σ_2 = ratio of standard deviations in population 1 and 2; δ = true and expected effect size (Cohen's d in population); $SPRT$ = sequential probability ratio test (Hajnal's t test) assuming $d = \delta$ and $\alpha = \beta = .05$. GS = group sequential design with four tests, assuming $d = \delta$ and $\alpha = \beta = .05$. SBF = sequential Bayes factor design with threshold 10, assuming r = $\sqrt{2}/2$, 1, $\sqrt{2}$ when δ = 0.2, 0.5, 0.8, respectively.

the variance ratio as long as the group sample sizes are balanced. Thus, our simulations show that for a balanced sampling scheme, Hajnal's *t* test is robust against violations of homoscedasticity assumptions.

The GS seems to be quite robust as well (there is virtually no effect in terms of efficiency), although its empirical Type 1 error rates slightly exceed the nominal level. The SBF design is quite robust when there is an effect (Type 2 errors), but there is a noticeable increase in Type 1 error rates in the case of unequal variances. Thus, SBFs seem to be affected by heteroscedasticity to a certain extent even when group sample sizes are balanced.

If sample sizes are unbalanced, Hajnal's *t* test and the SBFs are affected in quite the same manner. If there is homoscedasticity, error rates do not change, whereas their efficiency is lowered: Expected sample sizes of both tests increase notably. In the case of heteroscedasticity, however, both tests show poor Type 1 error rates when the sample with larger variance is smaller. This increase in error rates is not surprising, as the pooled variance estimate will seriously underestimate the true variance if the sample with larger variance is notably smaller than the other sample. This, in turn, will result in too large *t t* test and a high number of false-positive decisions. If the population with larger variance is overrepresented, on the other hand, both tests become more conservative and less efficient.

Interestingly, the GS is affected most seriously by an unbalanced design. Whereas Type 1 error rates are highly conservative when there is heteroscedasticity and the sample with small variance is larger, Type 2 error rates are inflated for all variance ratios. Hence, independent from heteroscedasticity, the GS design is strongly affected by unequal sample sizes.

Random Effects

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

In this simulation, a population effect size δ was randomly drawn from a normal distribution with $\mu_\delta = 0.2$, 0.5, 0.8 and $\sigma_\delta = 1$ in a first step. Subsequently, ran-

Table 4

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

	SPRT		GS		SBF	
Effect size	β'	ASN	β'	ASN	β'	ASN
$\delta \sim \mathcal{N}(0.2, 1)$	8.5	278	8.5	518	2.4	1042
$\delta \sim \mathcal{N}(0.5, 1)$	18.6	88	18.6	112	5.1	370
$\delta \sim \mathcal{N}(0.8, 1)$	23.6	46	23.5	52	7.0	198

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

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

Results. The empirical error rates and ASN can be obtained from Table 4. Not surprisingly, the error rates of Hajnal's t test and the GS are basically equivalent since they make the same assumption. However, as this assumption (fixed effect) is violated, the resulting error rates are seriously inflated. Hence, if the true effect size is in fact random rather than fixed or at least as large as expected, neither Hajnal's t test nor the GS can control the probability of a decision error.

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

Truncation Before a Decision

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

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

Table 5

Percentage of Type 1 and 2 Decision Errors and Expected Sample Size of Hajnal's t Test when Truncated at N_{NP}

δ	d	α' / β'	ASN	N_{NP}	% NP
0.0	0.2	6.9	800	1302	9.1
	0.5	6.7	134	210	10.6
	0.8	6.0	54	84	10.4
0.2	0.2	6.5	652	1302	8.7
0.5	0.5	6.4	112	210	9.9
0.8	0.8	5.7	48	84	10.5

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

thus be biased.

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

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

Empirical Example

In this section, we illustrate Hajnal's *t* test by applying it to a real data set. Following Schönbrodt et al. (2017), we chose open data from a replication of the *retrospective gambler's fallacy* (RGF) in the Many Labs Replication Project (Klein et al., 2014, https://osf.io/ydpbf/). The RGF, initially reported by Oppenheimer and Monin (2009), refers to people's false belief that seemingly rare outcomes are more likely to stem from a larger number of trials than seemingly common outcomes. In the experiment, participants are asked to imagine walking into a casino and observing a man rolling a die three times in a row. In the experimental condition, all dice show 6's, whereas in the control condition, two of the dice come up 6's while the third die comes up 3. Based on this scenario, participants are asked to indicate how many times they think the die had been rolled before they walked into the casino. In line with the theory, participants in the experimental group typically indicate a larger number of rolls than in the control condition. In the original study, Oppenheimer and Monin (2009) reported an effect size of Cohen's $d = 0.69$, 95% CI [0.16, 1.21]. In the replication study, the effect was reproduced with a total sample of $N = 5,942$ participants, Cohen's $d = 0.63$ [0.57, 0.68].

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

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

Discussion

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

When conceiving statistical inference as decision-making under uncertainty, error probabilities in statistical decisions must not be ignored, irrespective of the statistical framework used for making inferences (Lakens, 2016). Hence, employing test procedures and stopping rules that allow for error probability control is pivotal for the scientific endeavor. However, when applying statistical tests researchers also face practical constraints such as limited resources. This has led to a widespread neglect of statistical power and invited a number of questionable practices, which played their part in the development of the reproducibility crisis in psychology (Bakker, van Dijk, & Wicherts, 2012; Simmons et al., 2011). Thus, in order to improve current statistical practice, sensible and efficient alternatives are needed, for example, sequential methods. Although sequential hypothesis tests have been proposed to the field of psychology in the past, their application is still surprisingly scarce in experimental research (Botella et al., 2006; Lakens, 2014).

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

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

Second, we extended the comparison of SBFs and the GS design by Schönbrodt et al. (2017) and included the SPRT. We showed that the SPRT is not only more efficient than the GS but also than SBFs for a correctly specified hypothesis. However, it is not our intention to take a stance in the somewhat ideological quarrel between different schools of statistical inference. We merely point out the SPRT as an alternative to SBFs

[6] Note that this estimate of Cohen's d as well as the CI are based on the assumption of a fixed sample size and, thus, might be biased towards an overestimation of the true effect size. See the Discussion Section for details.

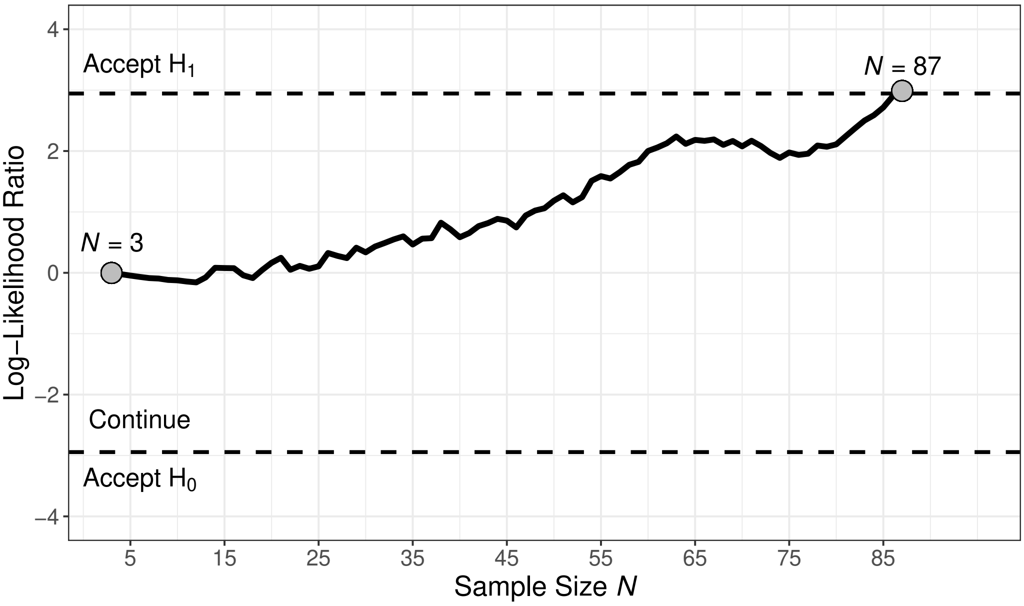

Figure 3. Development of the log-likelihood ratio for Hajnal's *t* test on the replication data of the retrospective gambler's fallacy (Klein et al., 2014). The test terminates sampling after $N = 87$ observations with a decision in favor of $\mathcal{H}_1$. The upper and lower dashed lines represent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

that (1) is more efficient when the alternative hypothesis corresponds to a point hypothesis and (2) allows for explicit control of error probabilities. If the psychological hypothesis of interest is in fact best represented by a prior distribution rather than a point mass, we endorse the use of a correspondingly specified likelihood ratio such as the Bayes factor implanted in SBFs. In the same vein, if the research goal is to quantify evidence and assess posterior probabilities, the SBFs (or generally, the Bayes factor) is the way to go. However, the standard SBF design does not allow for explicit control of error probabilities, which is a notable limitation. If error probability control is essential, the SPRT might constitute a better alternative.

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

Limitations

There are some possible limitations of our work that apply to sequential procedures in general, whereas others are specific for the test we promote in this article. First, some critics might object that the SPRT requires a precise specification of both the null and the alternative hypothesis (i.e., a precise prediction of the effect size). Ideally, this prediction follows from an underlying theory; however, it is frequently argued that researchers do not have realistic effect-size assumptions (Gelman & Carlin, 2014; Perugini et al., 2014). If there is no information in the literature such that an effect-size estimate could be based on a review or meta-analysis, this may indeed seem like a severe drawback. However, it is important to keep in mind that the effect-size assumption under $\mathcal{H}_1$ is not necessarily an attempt to guess the true effect that underlies the data. Alternatively, it can be seen as specification of an effect that the researcher "deem[s] worthy of detecting" (Schulz & Grimes, 2005, p. 1350). Thus, the need to specify a precise hypothesis should not be considered detrimental. After all, a hypothesis test is, by definition, the test of a prediction—why would we demand it to work without specifying one? NHST is an inglorious example of the critical consequences of employing a test without specifying a precise alternative to the null hypothesis.

It is true, however, that the SPRT will be less efficient or may lead to wrong decisions more often when the effect size is grossly under- or overspecified, respectively. At the cost of efficiency, the SBF design is more robust against such misspecifications to a certain extent. However, this does by no means free from the need to define a sensible statistical hypothesis: If the prior allocates undue mass to effect sizes that differ substantially from the true effect, the resulting test procedure will also perform poorly in terms of asymptotic error rates and efficiency. In sum, sensible hypothesis tests require reasonable and precise statistical hypotheses; the more precise a hypothesis, the more critical and efficient is its test (Stefan et al., 2019).

Second, the SPRT is an open procedure, that is, it requires sampling until a decision is made. It cannot be ruled out a priori that the data do not yield strong evidence in favor of any hypothesis such that the test goes on for thousands of observations. However, our results indicate that the risk of extremely large sample sizes in Hajnal's *t* test is small, although such events are possible in principle. Obviously, this is a potential risk in any open sequential design, SPRT and SBFs alike. Next to the GS design, there have been suggestions in the literature to modify sequential procedures such that they definitely terminate at or before a certain sample size N_{max} (T. W. Anderson, 1960; Armitage, 1957). However, as the comparison with the GS demonstrated, these restricted tests are not optimal, that is, they either are less powerful or come with higher ASN than open sequential designs (Wetherill, 1975).

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

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

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

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

Practical Recommendations

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

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

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

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

(1) Specify the statistical hypotheses (e.g., the to-be-detected minimal effect size d) and the desired upper bounds to the error probabilities of the test (α, β) before the sampling process. Do not alter these specifications during the sampling process in response to the data observed.

(2) Analyze the data in the sequence in which they have been sampled. This sequence must not be altered to obtain a specific result (e.g., by dropping unwanted observations). Observations may be added and analyzed in groups rather than separately. However, this may result in a decrease of error probabilities and, correspondingly, efficiency.

(3) Continue sampling as long as $\beta/(1 - \alpha) < LR < (1 - \beta)/\alpha$ and terminate as soon as this inequality is violated, resulting in a decision in favor of $\mathcal{H}_0$ if $LR \leq \beta/(1 - \alpha)$ or $\mathcal{H}_1$ if $LR \geq (1 - \beta)/\alpha$.

Conclusion

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

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

References

Anderson, K. (2014). gsDesign: Group sequential design. Retrieved from http://CRAN.R-project.org/package=gsDesign

Anderson, T. W. (1960). A modification of the sequential probability ratio test to reduce the sample size. *The Annals of Mathematical Statistics*, *31*, 165–197. doi:10.1214/aoms/1177705996

Anscombe, F. J. (1954). Fixed-sample-size analysis of sequential observations. *Biometrics*, *10*, 89–100. doi:10.2307/3001665

Armitage, P. (1947). Some sequential tests of student's hypothesis. *Supplement to the Journal of the Royal Statistical Society*, *9*, 250–263. doi:10.2307/2984117

Armitage, P. (1957). Restricted sequential procedures. *Biometrika*, *44*, 9–26. doi:10.2307/2333237

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., ... Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*, 108–119. doi:10.1002/per.1919

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

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

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

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

Botella, J., Ximénez, C., Revuelta, J., & Suero, M. (2006). Optimization of sample size in controlled experiments: The CLAST rule. *Behavior Research Methods*, *38*, 65–76. doi:10.3758/BF03192751

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

Clopper, C. J., & Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, *26*, 404–413. doi:10.2307/2331986

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cox, D. R. (1952). Sequential tests for composite hypotheses. *Mathematical Proceedings of the Cambridge Philosophical Society*, *48*, 290–299. doi:10.1017/S030500410002764X

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

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

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, *6*, 621. doi:10.3389/fpsyg.2015.00621

Emerson, S. S., & Fleming, T. R. (1990). Parameter estimation following group sequential hypothesis testing. *Biometrika*, *77*, 875–892. doi:10.2307/2337110

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program.

Behavior Research Methods, Instruments, & Computers, *28*, 1–11. doi:10.3758/BF03203630

Etz, A. (2018). Introduction to the Concept of Likelihood and Its Applications. *Advances in Methods and Practices in Psychological Science*, *1*, 60–69. doi:10.1177/2515245917744314

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

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149–1160. doi:10.3758/BRM.41.4.1149

Gelman, A. (2016). Commentary on "Crisis in science? Or crisis in statistics! Mixed messages in statistics with impact on science". *Journal of Statistical Research*, *48-50*, 11–12.

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

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences* (pp. 311–339). Hillsdale, NY: Erlbaum.

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

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on "Estimating the reproducibility of psychological science". *Science*, *351*, 1037–1037. doi:10.1126/science.aad7243

Goodman, S. N. (1993). P values, hypothesis tests, and likelihood: Implications for epidemiology of a neglected historical debate. *American Journal of Epidemiology*, *137*, 485–496. doi:10.1093/oxfordjournals.aje.a116700

Goodman, S. N. (2007). Stopping at nothing? Some dilemmas of data monitoring in clinical trials. *Annals of Internal Medicine*, *146*, 882–887. doi:10.7326/0003-4819-146-12-200706190-00010

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

Halsey, L. G., Curran-Everett, D., Vowler, S. L., & Drummond, G. B. (2015). The fickle p value generates irreproducible results. *Nature Methods*, *12*, 179–185. doi:10.1038/nmeth.3288

Jeffreys, H. (1935). Some tests of significance, treated by the theory of probability. *Mathematical Proceedings of the Cambridge Philosophical Society*, *31*, 203–222. doi:10.1017/S030500410001330X

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

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

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B. J., Bahník, Š., Bernstein, M. J., ... Nosek, B. A. (2014). Investigating variation in replicability: A 'many labs' replication project. *Social Psychology*, *45*, 142–152. doi:10.1027/1864-9335/a000178

Köllerström, J., & Wetherill, G. B. (1979). SPRT's for the normal correlation coefficient. *Journal of the American Statistical Association*, *74*, 815–821. doi:10.2307/2286405

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

Lakens, D. (2016). *Dance of the Bayes factors*. The 20% Statistician. Retrieved from http://daniellakens.blogspot.de/2016/07/dance-of-bayes-factors.html

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

Lee, H., & Fung, K. Y. (1980). A monte carlo study on the robustness of the two-sample sequential t-test. *Journal of Statistical Computation and Simulation*, *10*, 297–307. doi:10.1080/00949658008810377

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

Matthes, T. K. (1963). On the optimality of sequential probability ratio tests. *The Annals of Mathematical Statistics*, *34*, 18–21. doi:10.1214/aoms/1177704239

Matzke, D., Nieuwenhuis, S., van Rijn, H., Slagter, H. A., van der Molen, M. W., & Wagenmakers, E.-J. (2015). The effect of horizontal eye movements on free recall: A preregistered adversarial collaboration. *Journal of Experimental Psychology: General*, *144*, e1–e15. doi:10.1037/xge0000038

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist*, *70*, 487–498. doi:10.1037/a0039400

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

Morey, R. D., & Rouder, J. N. (2015). BayesFactor: Computation of Bayes factors for common designs. Retrieved from https://CRAN.R-project.org/package=BayesFactor

Morey, R. D., Rouder, J. N., Verhagen, J., & Wagenmakers, E.-J. (2014). Why hypothesis tests are essential for psychological science: A comment on Cumming (2014). *Psychological Science*, *25*, 1289–1290. doi:10.1177/0956797614525969

Mueller, P. S., Montori, V. M., Bassler, D., Koenig, B. A., & Guyatt, G. H. (2007). Ethical issues in stopping randomized trials early because of apparent benefit. *Annals of Internal Medicine*, *146*, 878–882. doi:10.7326/0003-4819-146-12-200706190-00009

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London*, *231*, 289–337. doi:10.1098/rsta.1933.0009

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, aac4716–aac4716. doi:10.1126/science.aac4716

Oppenheimer, D. M., & Monin, B. (2009). The retrospective gambler's fallacy: Unlikely events, constructing the past, and multiple universes. *Judgment and Decision Making*, *4*, 326–334.

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science*, *7*, 528–530. doi:10.1177/1745691612465253

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

Popper, K. R. (1968). *The logic of scientific discovery* (3d ed. (revised)). London: Hutchinson.

Proschan, M. A., Lan, K. K. G., & Wittes, J. T. (2006). *Statistical monitoring of clinical trials: A unified approach*. OCLC: ocm71228590. New York, NY: Springer.

Psychonomic Society. (2012). Psychonomic Society Statistical Guidelines. Retrieved from http://www.psychonomic.org/page/statisticalguideline

R Core Team. (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from http://www.R-project.org/

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

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225–237. doi:10.3758/PBR.16.2.225

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

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

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, *25*, 128–142. doi:10.3758/s13423-017-1230-y

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*, *22*, 322–339. doi:10.1037/met0000061

Schulz, K. F., & Grimes, D. A. (2005). Sample size calculations in randomised trials: Mandatory and mystical. *The Lancet*, *365*, 1348–1353. doi:10.1016/S0140-6736(05)61034-3

Sedlmeier, P. (1996). Jenseits des Signifikanztest-Rituals: Ergänzungen und Alternativen [Beyond the ritual of significance testing: Alternative and supplementary methods]. *Methods of Psychological Research Online*, *1*. Retrieved from http://www.pabst-publishers.de/mpr/

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105*, 309–316. doi:10.1037/0033-2909.105.2.309

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

Sobel, M., & Wald, A. (1949). A sequential decision procedure for choosing one of three hypotheses concerning the unknown mean of a normal distribution. *The Annals of Mathematical Statistics*, *20*, 502–522. doi:10.1214/aoms/1177729944

Stallard, N., Todd, S., & Whitehead, J. (2008). Estimation following selection of the largest of two normal means. *Journal of Statistical Planning and Inference*, *138*, 1629–1638. doi:10.1016/j.jspi.2007.05.045

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

Tryon, W. W. (2016). Replication is about effect size: Comment on Maxwell, Lau, and Howard (2015). *American Psychologist*, *71*, 236–237. doi:10.1037/a0040191

Ulrich, R., Miller, J., & Erdfelder, E. (2018). Effect size estimation from t-statistics in the presence of publication bias: A brief review of existing approaches with some extensions. *Zeitschrift für Psychologie*, *226*, 56–80. doi:10.1027/2151-2604/a000319

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

Wagenmakers, E.-J., Beek, T. F., Rotteveel, M., Gierholz, A., Matzke, D., Steingroever, H., ... Pinto, Y. (2015). Turning the hands of time again: A purely confirmatory replication study and a Bayesian analysis. *Frontiers in Psychology*, *6*. doi:10.3389/fpsyg.2015.00494

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

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

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

Wetherill, G. B. (1975). *Sequential methods in statistics* (2. ed.). London: Chapman and Hall.

Whitehead, J. (1986). On the bias of maximum likelihood estimation following a sequential test. *Biometrika*, *73*, 573–581. doi:10.1093/biomet/73.3.573

Wrinch, D., & Jeffreys, H. (1921). On certain fundamental principles of scientific inquiry. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *42*, 369–390. doi:10.1080/14786442108633773

Zellner, A., & Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. *Trabajos de Estadistica Y de Investigacion Operativa*, *31*, 585–603. doi:10.1007/BF02888369

Zhang, J. J., Blumenthal, G. M., He, K., Tang, S., Cortazar, P., & Sridhara, R. (2012). Overestimation of the effect size in group sequential trials. *Clinical Cancer Research*, *18*, 4872–4876. doi:10.1158/1078-0432.CCR-11-3118

(Appendix follows)

Table A1

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

		Hajnal's t test				Group Sequential Test				Sequential Bayes Factors		
		α = 1%		α = 5%		α = 1%		α = 5%				
d	r	β = 5%	β = 10%	β = 5%	β = 10%	β = 5%	β = 10%	β = 5%	β = 10%	BF = 5	BF = 10	BF = 30
							$\delta = 0$					
0.2	$\sqrt{2}/2$	888 (734,982,1722)	692 (560,764,1360)	832 (704,920,1512)	656 (544,732,1222)	1214 (964,1446,1928)	1006 (806,1208,1612)	1040 (1062,1062,1416)	834 (862,862,1148)	174 (136,186,396)	834 (594,844,1980)	8160 (5524,7878,19588)
0.5	1	146 (122,162,278)	116 (94,128,230)	138 (116,152,250)	110 (90,122,208)	198 (158,236,314)	166 (132,196,262)	170 (172,172,230)	138 (142,142,188)	92 (70,96,210)	426 (300,418,1010)	4152 (2782,3962,10202)
0.8	$\sqrt{2}$	60 (48,66,116)	48 (38,54,98)	56 (48,62,102)	46 (38,50,88)	80 (64,94,126)	68 (54,80,106)	68 (70,70,92)	56 (56,56,74)	48 (36,50,112)	216 (152,214,504)	2070 (1386,1986,4882)
							$\delta = 0.2$					
0.2	$\sqrt{2}/2$	998 870,1254,2072)	930 (812,1164,1890)	690 (592,880,1538)	626 (534,788,1350)	1278 (1446,1446,1928)	1150 (1208,1612,1612)	932 (1062,1062,1416)	808 (862,1148,1148)	236 (150,276,710)	770 (650,1078,1868)	1120 (994,1546,2518)
0.5	1	232 (160,286,592)	178 (118,218,474)	178 (134,212,412)	138 (102,160,320)	236 (236,314,314)	194 (196,262,262)	184 (172,230,230)	148 (142,188,188)	126 (74,128,390)	608 (450,848,1604)	1184 (1070,1636,2624)
0.8	$\sqrt{2}$	76 (54,84,182)	60 (42,66,144)	66 (50,76,144)	52 (38,58,110)	88 (94,94,126)	72 (80,80,106)	72 (70,92,92)	58 (56,74,74)	62 (38,60,178)	384 (204,484,1254)	1236 (1118,1696,2688)
							$\delta = 0.5$					
0.2	$\sqrt{2}/2$	276 (262,320,428)	272 (260,316,420)	194 (180,226,320)	190 (178,224,314)	496 (482,482,482)	436 (404,404,806)	378 (354,354,708)	328 (288,288,574)	100 (84,140,250)	132 (112,180,314)	172 (152,232,374)
0.5	1	168 (146,210,344)	158 (138,198,322)	120 (100,150,258)	110 (92,136,228)	208 (236,236,314)	186 (196,214,262)	150 (172,172,230)	132 (142,142,188)	88 (70,120,216)	134 (114,186,320)	176 (156,240,380)
0.8	$\sqrt{2}$	112 (86,142,266)	92 (70,116,224)	78 (60,96,180)	62 (48,76,144)	102 (94,126,126)	86 (80,106,106)	74 (70,92,92)	60 (56,74,74)	64 (44,82,170)	138 (120,192,324)	184 (164,248,404)
							$\delta = 0.8$					
0.2	$\sqrt{2}/2$	166 (160,184,226)	164 (158,182,222)	116 (112,130,166)	114 (110,128,162)	482 (482,482,482)	404 (404,404,404)	354 (354,354,354)	290 (288,288,288)	44 (36,60,108)	56 (48,74,124)	72 (64,94,148)
0.5	1	86 (80,102,148)	86 (78,100,144)	62 (54,72,112)	60 (54,72,110)	124 (158,158,158)	114 (132,132,196)	94 (116,116,172)	86 (94,94,142)	44 (36,60,106)	56 (48,74,126)	72 (64,94,150)
0.8	$\sqrt{2}$	72 (62,88,142)	68 (58,82,130)	50 (42,62,106)	48 (40,58,96)	82 (94,94,126)	74 (80,80,106)	60 (70,70,92)	52 (56,56,74)	42 (34,56,98)	58 (48,78,130)	74 (66,98,154)

Note. Depicted are expected total sample sizes ($n_1 + n_2$) and the 50th, 75th, and 95th quantile in parentheses. Note that the d and r metrics indicate fundamentally different effect-size expectations, although they are often assigned the same verbal labels for "small" ($d = 0.2$, $r = \sqrt{2}/2$), "medium" ($d = 0.5$, $r = 1$), and "large" effects ($d = 0.8$, $r = \sqrt{2}$). The Group Sequential Test comprised three interim and one final test. d = expected effect size according to $\mathcal{H}_1$ in Hajnal's t test and Group Sequential test (Cohen's d); r = scale parameter of Cauchy prior (= expected median absolute effect size) according to $\mathcal{H}_1$ in Sequential Bayes Factors; δ = true population effect size; BF = threshold Bayes factor.

Table A2

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

					$\delta = 0.2$			$\delta = 0.5$			$\delta = 0.8$		
Distribution	γ	s	k	True state	SPRT	GS	SBF	SPRT	GS	SBF	SPRT	GS	SBF
Normal		0.0	3.0	$\mathcal{H}_0$	832	1040	834	138	170	426	56	68	216
				$\mathcal{H}_1$	690	932	770	120	150	134	50	60	58
Mixture	.9	0.8	6.0	$\mathcal{H}_0$	834	1036	832	140	170	432	58	68	222
				$\mathcal{H}_1$	674	914	748	114	144	130	48	58	54
	.7	0.9	4.4	$\mathcal{H}_0$	830	1018	822	140	166	434	58	68	218
				$\mathcal{H}_1$	626	872	698	108	140	118	46	54	48
	.5	0.7	3.4	$\mathcal{H}_0$	794	994	732	138	162	384	58	66	204
				$\mathcal{H}_1$	562	832	612	96	134	104	42	52	46
Log-normal		6.2	116.9	$\mathcal{H}_0$	860	1040	886	146	172	454	60	70	232
				$\mathcal{H}_1$	642	892	682	94	132	96	36	48	36

Note. Depicted are expected total sample sizes ($n_1 + n_2$). The first two rows display results from the first simulation for normally distributed data, see Table 1, columns 3, 7, and 10. Number of repetitions per parameter combination: $k = 10,000$. γ = mixture probability; s = skewness; k = kurtosis; δ = true and expected effect size (Cohen's d in population); $SPRT$ = sequential probability ratio test (Hajnal's t test) assuming $d = \delta$ and $\alpha = \beta = .05$. GS = group sequential design with four tests, assuming $d = \delta$ and $\alpha = \beta = .05$. SBF = sequential Bayes factor design with threshold 10, assuming r = $\sqrt{2}/2$, 1, $\sqrt{2}$ when δ = 0.2, 0.5, 0.8, respectively; $\mathcal{H}_0$, $\mathcal{H}_1$ = true state underlying data generation.

Table A3

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

			$\delta = 0.2$			$\delta = 0.5$			$\delta = 0.8$		
N_1/N_2	σ_1/σ_2		SPRT	GS	SBF	SPRT	GS	SBF	SPRT	GS	SBF
1	1/4	$\mathcal{H}_0$	832	1028	797	137	166	410	56	67	204
		$\mathcal{H}_1$	683	924	738	116	148	129	49	58	54
	1	$\mathcal{H}_0$	834	1024	820	139	166	416	57	67	212
		$\mathcal{H}_1$	688	926	770	119	149	133	50	59	57
	4	$\mathcal{H}_0$	831	1023	784	137	167	408	56	67	210
		$\mathcal{H}_1$	687	928	744	115	149	128	49	58	53
1/3	1/4	$\mathcal{H}_0$	993	935	914	164	152	454	67	62	232
		$\mathcal{H}_1$	1471	1213	1871	240	195	282	99	77	112
	1	$\mathcal{H}_0$	1115	1027	1112	185	166	573	77	68	292
		$\mathcal{H}_1$	916	1025	1044	157	166	186	68	65	78
	4	$\mathcal{H}_0$	875	1025	507	152	164	286	65	65	164
		$\mathcal{H}_1$	533	792	321	95	127	81	42	49	39

Note. Depicted are expected total sample sizes ($n_1 + n_2$). Number of repetitions per parameter combination: $k = 10,000$. N_1/N_2 = ratio of sample sizes in group 1 and 2; σ_1/σ_2 = ratio of standard deviations in population 1 and 2; δ = true and expected effect size (Cohen's d in population); $SPRT$ = sequential probability ratio test (Hajnal's t test) assuming $d = \delta$ and $\alpha = \beta = .05$. GS = group sequential design with four tests, assuming $d = \delta$ and $\alpha = \beta = .05$. SBF = sequential Bayes factor design with threshold 10, assuming r = $\sqrt{2}/2$, 1, $\sqrt{2}$ when δ = 0.2, 0.5, 0.8, respectively; $\mathcal{H}_0$, $\mathcal{H}_1$ = true state underlying data generation.

Waldian t tests for accepting and rejecting the null hypothesis with controlled error probabilities

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

[1] University of Mannheim
[2] Philipps-Universität Marburg

Abstract

Rouder, Speckman, Sun, Morey, and Iverson (*Psychonomic Bulletin & Review*, 2009) proposed Bayesian t tests for accepting and rejecting the null hypothesis as an alternative to null-hypothesis significance testing (NHST). We endorse the necessity of statistical tests that allow for substantiated decisions not only against but also in favor of the null. However, a major drawback of Bayesian t tests is that error probabilities of statistical decisions remain uncontrolled. To remedy this problem, we propose a sequential probability ratio test that combines the Bayes factor proposed by Rouder et al. (2009) with decision criteria developed by Abraham Wald in 1947. We demonstrate by means of simulations that the corresponding sequential procedure, which we call *Waldian t test*, reliably controls decision error probabilities, with the nominal Type-1 and Type-2 error probabilities serving as upper bounds to the actual error rates. Moreover, Waldian t tests are easily implemented in practice. Finally, we critically discuss conventional criteria of interpreting Bayes factors as "moderate" or "strong" evidence for statistical hypotheses by showing that these criteria may imply error probabilities considerably larger than those researchers typically aim at.

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 observed data with respect to predictions from theories, that is, hypothesis testing (Morey, Rouder, Verhagen, & Wagenmakers, 2014). The statistical toolbox contains a vast number of different inference procedures, albeit not all equally suited for the purpose of hypothesis testing. Depending

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

This research was supported by a grant from the Deutsche Forschungsgemeinschaft (DFG, GRK 2277) to the Research Training Group "Statistical Modeling in Psychology" (SMiP).

Correspondence concerning this article should be addressed to Martin Schnuerch, Schloss, 68131 Mannheim, Germany. E-mail: martin.schnuerch@psychologie.uni-mannheim.de

on the researcher's intentions and the situation at hand, there are certain desiderata for a good statistical test. Ironically, *null-hypothesis significance testing* (NHST)—the dominant test procedure in psychology—hardly satisfies any of them (Wagenmakers, 2007) and has been criticized for decades as a theoretically unsound and flawed method (e.g., Bakan, 1966; Bredenkamp, 1972; Cohen, 1994; Gelman, 2016; Gigerenzer, 1993; Gigerenzer, 2004; Rozeboom, 1960). Yet, its shortcomings have only recently received broader attention and been acknowledged by the scientific community, as the replication crisis in psychology has fostered calls for a paradigm shift away from NHST (e.g., Cumming, 2014; Dienes, 2011).

Ten years ago, in what is now one of the most often-cited articles in *Psychonomic Bulletin & Review*, Rouder, Speckman, Sun, Morey, and Iverson (2009) proposed Bayesian t tests as a viable alternative to NHST. Building upon the Bayes factor, the method of choice for hypothesis testing and model selection in Bayesian statistics (Berger, 2006), Bayesian t tests quantify the relative evidence in the data for one hypothesis, typically referred to as *null hypothesis* ($\mathcal{H}_0$) vis-à-vis another, termed *alternative hypothesis* ($\mathcal{H}_1$). While $\mathcal{H}_0$ usually is a simple point hypothesis on a parameter similarly as in NHST (e.g., $\mathcal{H}_0$: $\delta = 0$ when testing whether a difference in means is zero), $\mathcal{H}_1$ typically is a distributional hypothesis that specifies a prior distribution over the range of possible parameter values, for example, a Cauchy distribution ($\mathcal{H}_1$: $\delta \sim$ Cauchy). The Bayes factor quantifies the relative evidence for the two hypotheses of interest and is defined as the multiplicative updating factor for transforming prior beliefs for the competing hypotheses to posterior beliefs. Essentially, the Bayes factor prefers the model that predicted the data best by considering how well the observed data match with the hypotheses of interest, thus reflecting a central goal of statistical inference from a Bayesian perspective (Morey, Romeijn, & Rouder, 2016; Rouder, Morey, Verhagen, Province, & Wagenmakers, 2016).

Bayesian t tests (i.e., applications of the the Bayes factor with suitable $\mathcal{H}_0$ and $\mathcal{H}_1$) provide a remedy for several of the critical problems associated with NHST (Wagenmakers, 2007). Most importantly, unlike p values, the Bayes factor can provide evidence *in favor* of the null hypothesis (Kass & Raftery, 1995; Rouder et al., 2009). In fact, Bayes factors satisfy many of the desiderata for a good statistical procedure and thus have attracted notable attention in the field of psychology. As a rough proxy for the increasing popularity of Bayesian hypothesis tests, Figure 1 displays the number of peer-reviewed articles registered on PsycINFO featuring the keyword *Bayes factor*, published over the last ten years. Both the number of methodological articles as well as empirical publications using the Bayes factor have increased substantially.

The growing awareness for the limitations of NHST and the corresponding shift toward other, better justified statistical methods is a most welcome development. We endorse the use of inferential procedures such as Bayesian t tests that put emphasis on the explicit specification of both the null and the alternative hypothesis and allow for substantiated inference in favor of either hypothesis (unlike NHST). As outlined by Rouder et al. (2009), this is particularly useful if a strong psychological theory predicts a point null hypothesis, contrasted against an alternative hypothesis that allows for deviations from the null as formalized by a suitable prior distribution. However, as detailed below, the default Bayesian t test does not satisfy all relevant desiderata: As a continuous measure of evidence, the Bayes factor does not constitute a natural basis for binary statistical decisions—such as accepting or rejecting the null hypothesis—with controlled error probabilities (Schnuerch &

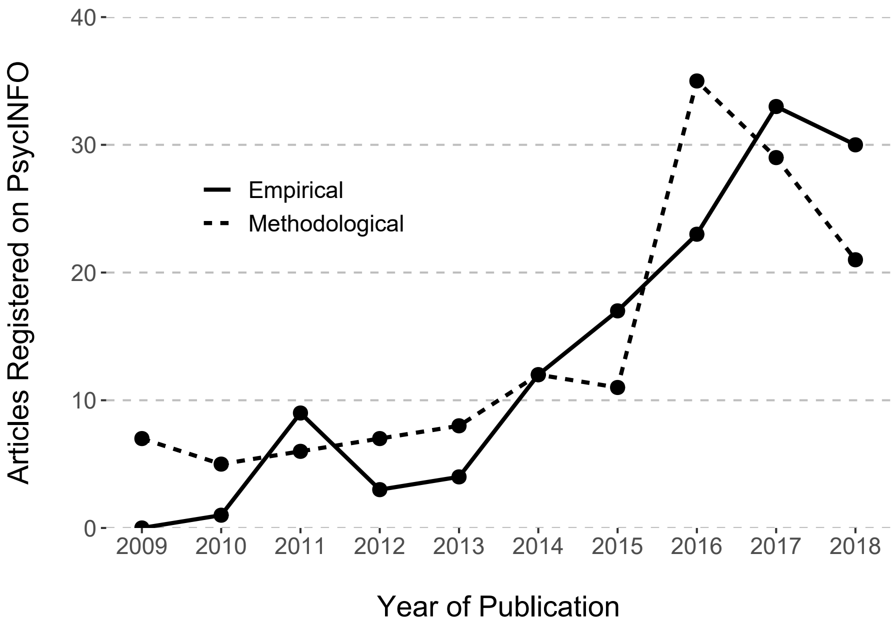

Figure 1. Number of articles containing the keyword *Bayes factor* registered on PsycINFO by year of publication. Only peer-reviewed journal articles were considered.

Erdfelder, 2019). This is a notable limitation for applications in which statistical decisions are required.

As a remedy, we propose a simple extension for Bayesian t tests when the aim is, in fact, to accept or reject a null hypothesis. This extension is based on Wald's (1947) sequential probability ratio test and thus called *Waldian t test* in what follows. Waldian t tests control long-term error rates in the classical sense while using exactly the same specifications of $\mathcal{H}_0$ and $\mathcal{H}_1$ as Bayesian t tests, thus preserving the interpretation of the corresponding Bayes factor. We argue that Waldian t tests satisfy all relevant desiderata of a hypothesis-testing procedure when a statistical decision in favor or against a point null hypothesis is required and, thus, represent a useful extension of existing inferential procedures.

Properties of a Good Statistical Procedure

Researchers and even statisticians hold different opinions as to which properties a good statistical inference procedure should possess (e.g., Dienes, 2011; Wagenmakers, 2007). Herein, we focus on three desiderata we deem particularly important (see Berger & Bayarri, 2004; Neyman, 1977; Royall, 1997)

First, a sensible statistical test must be able to convey support for any of the hypotheses tested, that is, both the null and the alternative hypothesis. It is well known that NHST fails to satisfy this desideratum (Wagenmakers et al., 2017). In NHST, only the null hypothesis is specified and inference is based on the p value, denoting the conditional probability of the observed or more extreme data, given that the null hypothesis holds. The logic of this test has been termed "Fisher's disjunction" (Rouder, Morey, Verhagen, et al., 2016): A small p indicates that either a rare event has been observed or the null hypothesis is wrong.

Thus, given a small p value, the (unspecified) alternative hypothesis is accepted and we are inclined to conclude that it is true, rather than the null hypothesis. Not only is this reasoning flawed—a procedure cannot measure evidence for a hypothesis that has not been specified—also, the interpretation of large p values as evidence for the null hypothesis is inadmissable as well. If the null hypothesis is false, the p value indeed converges to zero as the sample size increases toward infinity (Rouder et al., 2009). However, if the null is true, the p value does not converge to any fixed value but rather follows a uniform distribution in the unit interval for all sample sizes. Thus, the p value can only measure evidence *against*, but never *in favor of* the null hypothesis (Rouder et al., 2009; Wagenmakers, 2007).

Bayesian t tests, in contrast, satisfy the requirement of being able to provide support for any of the two hypotheses. Inference in Bayesian t tests is based on a quantity typically attributed to the works of Sir Harold Jeffreys (Etz & Wagenmakers, 2017). This quantity represents the factor by which the relative belief in two competing hypotheses *before* seeing the data (i.e., the prior odds) is updated in order to arrive at the relative belief *after* seeing the data (i.e., the posterior odds). The multiplicative updating factor is the *Bayes factor* (Jeffreys, 1961; Kass & Raftery, 1995):

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

The term $f(\text{data}|\mathcal{H}_i)$ denotes the marginal likelihood of $\mathcal{H}_i$, that is, the probability (density) of the observed data under hypothesis i (Rouder et al., 2009). This formula follows directly from Bayes' rule. It clearly demonstrates that the factor by which subjective belief is updated is, in fact, the relative accuracy of the two hypotheses in predicting the observed data (Rouder & Morey, 2017). Since predictive accuracy denotes the relative evidence in the data for each hypothesis (Morey et al., 2016; Royall, 1997), Bayesian t tests can measure support for any of the specified hypotheses, vis-à-vis the other.

Second, a good statistical procedure should provide informative results in an efficient way, that is, with sample sizes as small as possible. In NHST, an informative result typically means a "significant" result, that is, $p \leq \alpha$, where α denotes the significance level defined a priori (typically, $\alpha = .05$). So what happens when researchers start with a cost-efficient sample of, say, 40 participants and are left with $p > \alpha$? As outlined before, this result is inconclusive as it neither provides evidence against nor in favor of the null hypothesis. Can this ambiguity be removed by successively increasing the sample, followed by further statistical tests? The answer is negative: If the null hypothesis is true, the statistical test will keep its pre-specified α level only when inspecting the data once. If multiple tests are conducted repeatedly while stopping only when an "informative result" is obtained ($p \leq \alpha$) and increasing the sample size otherwise, the probability of the p value falling below α at some point approaches one—even when the null hypothesis is true (Armitage, McPherson, & Rowe, 1969). This questionable research practice represents a form of p hacking called *data peeking* (e.g., Erdfelder & Heck, in press; Simonsohn, Nelson, & Simmons, 2014).

Bayesian t tests, in constrast, satisfy the efficiency requirement. According to the *likelihood principle* (Berger & Wolpert, 1988), the evidential interpretation of a ratio of likelihoods only depends on the data observed, irrespective of how the data came about (e.g., which stopping rule was used for sample-size determination). As a result, Bayes factors can be computed repeatedly during the sampling process without altering their interpretation, allowing for optionally stopping or increasing the sample size at any point (Edwards, Lindman, & Savage, 1963; Lindley, 1957; Rouder, 2014). The flexibility of sequential testing makes Bayesian t tests more efficient than their NHST counterpart: Whenever the available evidence appears ambiguous, one can add more observations until sufficient evidence has been collected. In the same vein, sampling can be terminated whenever the evidence is deemed sufficient (Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, 2017).

To illustrate the advantages of optional stopping, consider the example depicted in Figure 2. It displays the development of a Bayes factor with standard Cauchy prior for a two-sample t test, denoting how well the data were predicted by the alternative hypothesis relative to the null hypothesis (BF_{10}). The Bayes factor was computed using the R package *BayesFactor* (Morey & Rouder, 2015), and the underlying data were simulated from two normal distributions with a standardized mean difference of $\delta = 0.50$ (i.e., the population value of Cohen's d). Assume two experimenters, A and B, decided to terminate sampling after 70 observations and analyze the data using an NHST with $\alpha = .05$ (Experimenter A) and a standard Bayesian t test (Experimenter B). As Figure 2 shows, Experimenter A finds the data to be inconclusive, $p > \alpha$, indicating that the null cannot be rejected. If A had decided to sample 200 observations, however, the result would have been different. As $p < \alpha$, she would have rejected the null hypothesis. Because she decided a priori to inspect the data after 70 observations, however, she cannot continue with subsequent data collection and testing, and is thus left with an inconclusive result. After 70 observations, the conclusion of Experimenter B agrees with that of Experimenter A: The resulting Bayes factor $BF_{10} = 1.08$ suggests that the evidence in the data favors none of the hypotheses over the other. However, unlike Experimenter A, Experimenter B can just continue the experiment until the data show more unequivocal evidence in favor of one of the hypotheses (e.g., after 164 observations, when $BF_{10} = 10.97$, or after 200 observations, when $BF_{10} = 15.46$).

As for the third desideratum, a good statistical procedure should have good long-run properties (Sanborn & Hills, 2014). If applied correctly, classical t tests control the long-run frequency of incorrect rejections of the null hypothesis. Without consideration of a specific alternative, however, there is no means to control the probability of incorrectly accepting the null if the alternative is true. In fact, the normative standard of classical hypothesis testing, that is, the procedure for statistical decision making introduced by Neyman and Pearson (1933), allows for control of both the Type-1 and Type-2 error probabilities by means of an a priori power analysis (Cohen, 1988). However, NHST as the de facto standard in psychological research ignores the alternative and, thus, the issue of statistical power (Gigerenzer, 1993, 2004).

In the standard Bayesian t tests, there is no means to control long-run frequencies of incorrect decisions, either. Instead, the Bayes factor is defined as a continuous measure of evidence for informing one's subjective beliefs about the hypotheses. Accordingly, thresholds for the interpretation of Bayes factors (e.g., $BF > 3$ or $BF > 10$ for "weak" or "strong" evidence, respectively) are merely based on conventions (e.g., Jeffreys, 1961; Wagenmakers et al., 2018). There are currently no normative, theoretically derived thresholds for the Bayes factor which, if exceeded, mandate a decision to reject or accept a hypothesis. It has frequently been argued that communicating evidence and posterior probabilities (or odds) on a continuous scale, rather than making decisions, truly reflects the aim of statistical

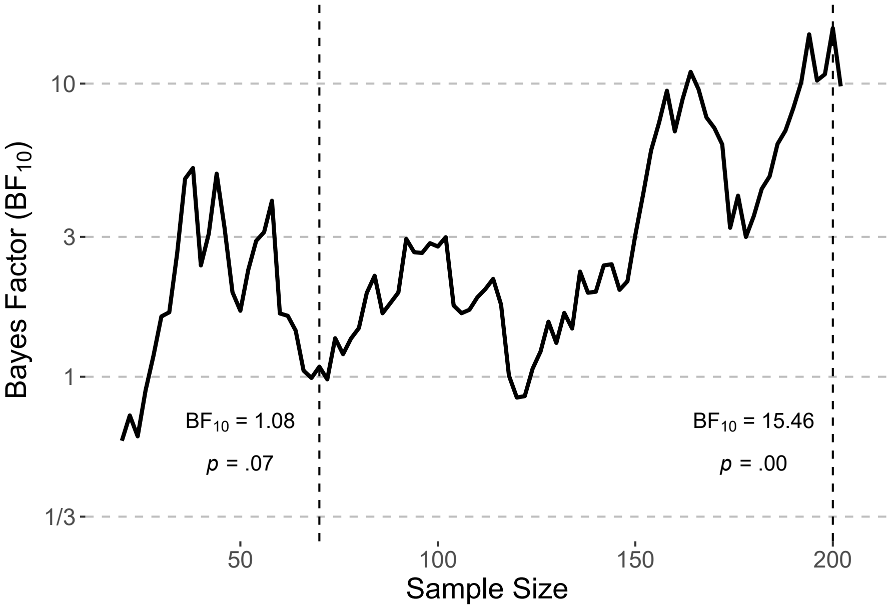

Figure 2. Development of the Bayes factor (BF_{10}) with standard Cauchy prior for a two-sample t test. True effect size $\delta = 0.50$. Dashed lines represent inspections after 70 and 200 observations, respectively.

inference (e.g., Dienes, 2011; Edwards et al., 1963; Morey et al., 2016; Rouder, Morey, & Wagenmakers, 2016; Rozeboom, 1960). In practice, however, researchers might well be inclined to dichotomize the continuous Bayes factor into regions of acceptance or rejection (Jeon & De Boeck, 2017). In fact, as we will outline below, many situations compel researchers to make decisions. In these situations, the long-run rates of incorrect decisions of a statistical procedure have to be considered (Sanborn & Hills, 2014). Thus, for these situations standard Bayesian t tests fail to satisfy a relevant desideratum.

Subjective Belief versus Error Control

From a strict Bayesian perspective, probability denotes subjective belief and Bayesian t tests represent a principled way to update that belief in light of the data (Morey et al., 2016). From this point of view, statistical inference should be about inductive probabilistic statements and one might consider long-run error rates irrelevant, since they refer to conditional probabilities assuming that one of the hypotheses is true (e.g., Wagenmakers & Gronau, 2018). Rouder (2014) states 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).

Unless scientists actually adhere to this principle, however, and completely refrain from binary decisions, the properties of the Bayesian inference procedure in terms of decision accuracy are highly relevant but unknown (Jeon & De Boeck, 2017). Heuristic taxonomies for interpreting the strength of evidence conveyed by the Bayes factor (Jeffreys, 1961; Lee & Wagenmakers, 2013), for example, invite practitioners to translate the Bayes factor into a discrete decision. If a Bayes factor greater 10 indicates "strong evidence" according to a widely used convention (Jeffreys, 1961), doesn't this imply that one can safely accept the respective hypothesis upon observing a Bayes factor larger than 10? However, how "safe" in terms of error rates would such a decision be? Since many researchers make use of these conventional taxonomies in applied Bayesian statistics, we suspect that most scientists aim at categorizing empirical evidence as either supporting or contradicting specific hypotheses.

In fact, in many contexts of research, both basic and applied, statistical inference necessitates a binary decision, namely, to accept or reject a hypothesis in the face of empirical data. Whenever specific actions are taken depending on the outcome of a statistical test, the inference becomes a decision between discrete options. Think of a clinical psychologist, for example, who has to decide whether or not to implement a new therapy. The decision will be made depending on whether or not the hypothesis is accepted that the new therapy is better than the old one. In the same vein, experimental psychologists might conduct a pilot study to test a specific hypothesis and decide to continue this line of research depending on whether or not the pilot study leads to (preliminary) acceptance of the hypothesis. In all of these cases, discrete decisions based on continuous statistical evidence cannot be avoided, which in turn implies that researchers should both quantify and minimize the error probabilities in these decisions.

Misusing Bayesian t tests as a decision-making tool bears the undetermined risk of producing high rates of false-positive results or underpowered studies —much like what we have witnessed in the context of NHST (Sanborn & Hills, 2014). Therefore, we emphasize the importance of considering frequentist concepts such as error rates and statistical power also in the context of Bayesian hypothesis testing.

This plea is not particularly new. A prominent Bayesian, James Berger, has long argued that "statisticians should readily use both Bayesian and frequentist ideas" (Berger & Bayarri, 2004, p.58). To unify the different schools of thought, Berger, Brown, and Wolpert (1994) introduced a strategy in which error rates from a frequentist perspective are equivalent to Bayesian posterior error probabilities (see also Bayarri, Benjamin, Berger, & Sellke, 2016; Berger, Boukai, & Wang, 1997; Berger, Boukai, & Wang, 1999). Their focus, however, was on post-experimental error probabilities conditioned on the observed data, which is referred to as *conditional* frequentist perspective (Kiefer, 1977). The focus of our present article, in contrast, is on promoting a specific design for Bayesian t tests such that the procedure does not exceed some predefined error probabilities. This is termed *unconditional* frequentist perspective, which corresponds to the Neyman-Pearson-Wald notion of error rates as stable properties of the test procedure (Royall, 1997).

Proper error control in this sense is about designing the test procedure such that one can control the probability of incorrect decisions, conditional on the hyptheses tested being true. Building on previous work by Wald (1947) and Berger et al. (1999), we argue that there is a simple means to control error probabilities of statistical decisions in Bayesian t tests: the sequential probability ratio test (SPRT). The SPRT is a sequential test procedure based on a likelihood ratio for which termination criteria are computed by simple formulae that satisfy prespecified error probabilities (α, β). We will show that this test is easily combined with a Bayesian t test for any proper prior distribution, and that it allows for control of decision error probabilities while maintaining the Bayesian specification of the models and priors for $\mathcal{H}_0$ and $\mathcal{H}_1$ as well as the interpretation of the Bayes factor. In that, we believe that the procedure we suggest unifies beneficial properties from both worlds.

Error Control for Bayesian t Tests

Bayesian t Tests

Defined as the ratio of posterior odds to prior odds, the Bayes factor in the Bayesian t tests captures all the information in the data about the hypotheses tested (see Equation 1). Let $x = (x_1, ..., x_n)$ denote a sample of n observations assumed to be distributed as $X \sim f(x|\theta)$, where θ is the parameter vector defining the probability density function $f(.)$. For hypothesis $\mathcal{H}_i$, $i \in \{0, 1\}$, the marginal likelihood M_i of the observed data is given by the integral

$$M_i = \int_{\theta \in \Theta_i} f(x|\theta, \mathcal{H}_i)\ \pi(\theta|\mathcal{H}_i)\ d\theta, \tag{2}$$

where Θ_i is the parameter space defined by hypothesis i and $\pi(.)$ denotes the prior distribution on the parameters θ over this space. As Equation 2 shows, M_i denotes a weighted average likelihood of the observed data under hypothesis i. The Bayes factor for the test of an alternative $\mathcal{H}_1$ against a null hypothesis $\mathcal{H}_0$ is simply the ratio of marginal likelihoods:

$$BF_{10} = \frac{M_1}{M_0}. \tag{3}$$

Being a likelihood ratio, the Bayes factor denotes the evidence provided by the data x for one statistical model relative to the other (Kass & Raftery, 1995). In case of Bayesian t tests, the Bayes factor can be interpreted as a test between two different prior distributions on the parameters,

$$\mathcal{H}_i : \theta \sim \pi(\theta|\mathcal{H}_i). \tag{4}$$

This highlights the relevance of the choice of prior distributions when computing and interpreting a Bayes factor. Essentially, one should carefully define these distributions for each hypothesis according to prior convictions or expectations one might hold about the unknown parameters (Rouder et al., 2009).

When defining a prior distribution for the t test, we are faced with the problem that we want to test hypotheses about population means or mean differences of normally distributed random variables without knowing the scale of the dependent variable (i.e., the population variance σ^2). Consider the common two-sample scenario: Let X and Y denote observations from two groups, modeled as

$$X \sim \mathcal{N}(\mu + \frac{\delta\sigma}{2}, \sigma^2) \tag{5}$$

and

$$Y \sim \mathcal{N}(\mu - \frac{\delta\sigma}{2}, \sigma^2), \tag{6}$$

with μ denoting the grand mean and δ the standardized effect size (i.e., Cohen's d; Cohen, 1988). The prior distributions on the population parameters δ, μ, and σ^2 suggested by Rouder et al. (2009) for Bayesian t tests are commonly referred to as *JZS prior*, because they are based on prior specifications by Jeffreys (1961) and Zellner and Siow (1980). The priors on μ and σ^2 are non-informative and identical under both hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$. Thus, the effect of these priors on the resulting Bayes factor is negligible, implying that the statistical hypotheses actually tested by a Bayesian t test are fully defined by the priors on δ (Rouder et al., 2009).

Under the null hypothesis, the prior on δ is a point mass corresponding to the classical null hypothesis in NHST that the group means are identical, $\mathcal{H}_0$: $\delta = 0$. Under the alternative, in contrast, the prior is a Cauchy distribution, that is, $\mathcal{H}_1$: $\delta \sim$ Cauchy. The Cauchy is a heavy-tailed distribution defined over the entire real line, centered at zero. Its shape is defined by a scale parameter r, such that 50% of its weight are assigned to values in the interval $[-r, r]$. Thus, if a smaller value of the scale parameter r is chosen, the prior distribution assigns more probability mass to smaller effect sizes around zero (see Figure 3). For $r = 1$, the Cauchy is a t distribution with one degree of freedom.

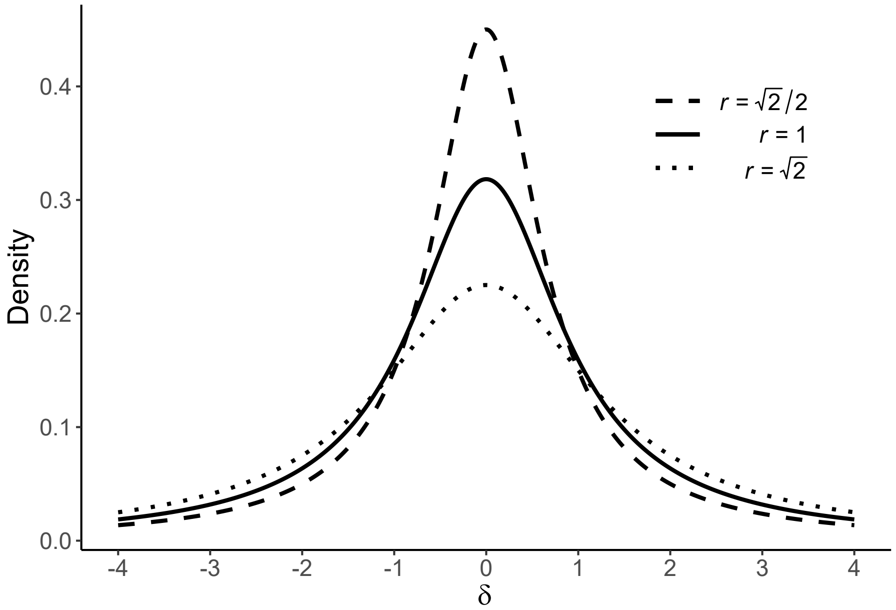

Figure 3. Shape of the Cauchy prior distribution for the standardized effect size δ (i.e., the population value of Cohen's d) in Bayesian t tests for different specifications of the scale parameter r.

The justification for the choice of the Cauchy distribution as a prior under the alternative hypothesis is primarily based on its favorable mathematical properties (for details, see Ly, Verhagen, & Wagenmakers, 2016). However, it also has a reasonable psychological interpretation and practical appeal: Given a theoretically motivated null hypothesis that two group means are expected to be identical, there might be no reason to assume a certain fixed effect size under the alternative. For such a case, the Cauchy might be interpreted as a representation of the distribution of possible non-zero effect sizes in the field when the predicted point hypothesis does not hold, specifically, when the true effect is randomly distributed around the effect predicted by the null. The Cauchy puts emphasis on smaller effect sizes around the null. At the same time, there is also substantial weight on large effect sizes, which are notably less frequent, albeit possible.

To summarize, Bayesian t tests are a theoretically sound method to quantify statistical evidence for or against a null vis-à-vis an alternative hypothesis in the common t-test situation. If the aim is to communicate this evidence on a continuous (Bayes factor) scale or employ it to update prior beliefs about the plausibility of the specified statistical models, there is no need for an extension of the method. However, when used as a decision-making tool for accepting or rejecting the null hypothesis, it is necessary to consider long-run properties. In the following section, we outline a simple statistical approach to control error probabilities in Bayesian t tests.

Waldian t Tests

Statistical procedures that are concerned with error probability control, such as Neyman-Pearson tests, are typically bound to the specification of so-called *simple* hypotheses. A hypothesis is simple if all parameter values are known or specified by the hypothesis. If a hypothesis is not simple, it is composite (Wald, 1947).

In default Bayesian t tests, a simple null hypothesis on the parameter of interest ($\delta = 0$) is tested against a composite alternative ($\delta \sim$ Cauchy). The problem is to define a statistical procedure for the Bayesian test such that the following requirements are satisfied:

$$P(\text{accept } \mathcal{H}_i|\mathcal{H}_i) = \begin{cases} 1-\alpha & (i=0) \\ 1-\beta & (i=1) \end{cases}, \tag{7}$$

where $P(\text{accept } \mathcal{H}_i|\mathcal{H}_i)$ denotes the conditional probability to accept hypothesis i when it is true, thus defining α and β as the Type-1 and Type-2 error probabilities, respectively. A solution for this problem has been outlined by Wald (1947) and discussed by Berger et al. (1999).

For a simple null hypothesis defined by a point in the parameter space (i.e., $\delta = 0$ in the t test), the probability of a Type-1 error has a fixed value α for any given test procedure. In contrast, for a composite alternative hypothesis $\mathcal{H}_1$, the probability of a Type-2 error depends on the specific parameter value δ under the alternative hypothesis and is a single-valued function of δ defined over the parameter space Δ_i under hypothesis i:

$$P(\text{reject } \mathcal{H}_1|\delta) = \beta(\delta),\ \forall \delta \in \Delta_1. \tag{8}$$

To take into account that not all values of δ are equally plausible, we can specify a non-negative weight function, $\omega_1(\delta)$, that integrates to one,

$$\int_{\delta \in \Delta_1} \omega_1(\delta)\ d\delta = 1. \tag{9}$$

Then, error control satisfying the requirements given in Equation 7 is achieved if we can define a test procedure such that the weighted average of $\beta(\delta)$ is equal to the desired Type-2 error probability,

$$P(\text{reject } \mathcal{H}_1|\mathcal{H}_1) = \int_{\delta \in \Delta_1} \beta(\delta)\ \omega_1(\delta)\ d(\delta) = \beta. \tag{10}$$

We further note that, by marginalizing over the free parameter δ weighted by $\omega_1(\delta)$ in Equation 10, the test of the composite hypothesis $\mathcal{H}_1$ on the distribution of the unknown parameter δ is in fact equivalent to the test of a simple hypothesis $\mathcal{H}_1^*$ on the probability distribution of the data,

$$\mathcal{H}_1^* :\ x \sim f_1(x) = \int_{\delta \in \Delta_1} f(x|\delta)\ \omega_1(\delta)\ d(\delta), \tag{11}$$

thus reducing the problem to a test of a simple null hypothesis against a simple alternative (Berger et al., 1997, 1999).

The Cauchy prior in the Bayesian t tests is a proper weight function as defined in Equation 9. Thus, the corresponding Bayes factor can be seen as a likelihood ratio for two simple hypotheses (see Equation 11). Consequently, a likelihood-ratio test procedure for the Bayes factor with Type-1 and Type-2 error probabilities α and β would satisfy the requirements for controlling error rates as specified in Equation 7.

Such a test procedure is given by the sequential probability ratio test (Wald, 1945, 1947). The SPRT is a sequential procedure for a test between two simple hypotheses. For this scenario, it has been proven to be the most efficient test. In other words, for given error probabilities α and β there is no alternative test that requires fewer observations on average (Wald & Wolfowitz, 1948; Wetherill, 1975).

The general procedure of the SPRT requires computation of the likelihood ratio LR for the observed data under the two hypotheses after every single observation, followed by a decision in line with the following three decision rules:

$$\begin{aligned}
&\text{1) Accept } \mathcal{H}_1 \text{ and reject } \mathcal{H}_0 \text{ when } LR \geq A;\\
&\text{2) Accept } \mathcal{H}_0 \text{ and reject } \mathcal{H}_1 \text{ when } LR \leq B;\\
&\text{3) Sample a new independent observation when } B < LR < A.
\end{aligned} \tag{12}$$

By Rule 1, any observed sample that leads to the acceptance of $\mathcal{H}_1$ is at least A times as likely under the alternative as under the null hypothesis. This implies, in turn, that the long-term probability of accepting $\mathcal{H}_1$ with this procedure is at least A times larger if $\mathcal{H}_1$ is in fact true (= correct acceptance) than if $\mathcal{H}_0$ is true (= Type-1 error), that is, $1 - \beta \geq A\alpha$. Following the same reasoning for the lower threshold specified in Rule 2, we see that $\beta \leq B(1 - \alpha)$. Rewriting these inequalities and replacing them by equalities provides definitions for the threshold values such that the error probability requirements specified in Equation 7 will be satisfied approximately (Wald, 1947):

$$A = \frac{1 - \beta}{\alpha} \tag{13}$$

and

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

Note that replacing the inequalities by equalities merely renders the procedure more conservative, that is, the nominal error probabilities α and β will be upper bounds to the actual error probabilities achieved by inserting Equations 13 and 14 in the sequential decision procedure (Equation 12).

As outlined above, Bayesian t tests can be conceptualized as a test of two simple hypotheses. Thus, the general logic of the SPRT applies. To control decision error probabilities, the Bayes factor can be computed sequentially until it reaches (or crosses) the upper or lower threshold A or B, respectively. As shown above, the long-term error rates of this procedure will approximate (but never exceed) α and β if researchers strictly follow the procedure defined in Equation 12 in combination with threshold values defined by Equations 13 and 14 (Wald, 1947).

This sequential design for Bayesian t tests, which we call Waldian t tests, has fully Bayesian and frequentist justification (Berger et al., 1999): Since the interpretation of the Bayes factor is unaffected by optional stopping (Rouder, 2014; but see Sanborn et al., 2014; Yu, Sprenger, Thomas, & Dougherty, 2014), the specific value of the Bayes factor obtained after reaching the upper or lower threshold A or B, respectively, in a Waldian t test preserves its fully Bayesian interpretation. At the same time, because threshold values based on α and β are defined for the likelihood ratio, it controls the probabilities of decision errors conditional on the specified hypotheses. Thereby, Waldian t tests combine the beneficial properties of Bayesian and classical Neyman-Pearson-Wald hypothesis testing, thus satisfying the three desiderata discussed above.

Simulation of Waldian t Tests

The properties of the SPRT have been derived and proven analytically (Wald, 1945, 1947; Wald & Wolfowitz, 1948). As outlined above, however, the nominal error probabilities used to define the decision thresholds A and B are upper bounds to the actual error rates: At the time of termination, the likelihood ratio (or, in case of a Waldian t test, the Bayes factor) will almost always exceed rather than be equal to one of the boundary values (so-called *overshoot*). When there is substantive overshoot, the resulting error rates of the sequential test procedure will undercut the nominal values (Berger et al., 1999; Wald, 1947). To examine the extent to which empirical error rates deviate from the nominal rates, we simulated the proposed Waldian t tests. Simulations have "a tangible, experimental feel" and are thus well suited to demonstrate the properties of an analytically justified approximate method to psychologists (Rouder, 2014, p. 303). All simulations and analysis were performed in R (R Core Team, 2019). Reproducible scripts and all simulated data are available at the Open Science Framework (https://osf.io/z5vsy/?view_only=63252a8bd2374b8297bf8bb9fdd124 42).

Design

We simulated the Waldian t tests based on default Bayesian t tests for two independent samples as proposed by Rouder et al. (2009). We chose the default test for the sake of demonstration. Note, however, that Waldian t tests provide a general framework that applies to any choice of proper priors, not just those proposed by Rouder et al. (2009). The statistical hypotheses tested are $\mathcal{H}_0$: $\delta = 0$ and $\mathcal{H}_1$: $\delta \sim \text{Cauchy}(r)$. Under $\mathcal{H}_0$, we simulated random data from two normal distributions with means $\mu_1 = \mu_2 = 0$ and common standard deviation $\sigma = 1$. Under $\mathcal{H}_1$, the true effect size δ was randomly drawn from the specified Cauchy distribution with the scale parameter r to generate data from two normal distributions with group means $\mu_1 = 0$ and $\mu_2 = \delta$.

Starting at an initial sample size of $n = 2$ per group, the Bayes factor was computed using the R package *BayesFactor* (Morey & Rouder, 2015). The sample was then increased stepwise by $+1$ in each group until the Bayes factor BF_{10} reached one of the boundary values $A = (1 - \beta)/\alpha$ or $B = \beta/(1 - \alpha)$. As soon as a threshold was reached, sampling was terminated and the respective hypothesis was accepted.

We systematically varied the scale parameter of the Cauchy distribution according to the default values provided in the BayesFactor package, that is, $r \in \{\sqrt{2}/2;\ 1;\ \sqrt{2}\}$.

Additionally, different combinations of nominal error probabilities of the Waldian t tests were simulated with $\alpha \in \{.01;\ .05;\ .10\}$ and $\beta \in \{.05;\ .10;\ .20\}$. For each parameter combination, $20,000$ replications were simulated.

Results

The simulation results are depicted in Figure 4. It displays the empirical error rates (i.e., the proportion of simulations with decisions in favor of the wrong statistical model) with 95% Clopper-Pearson exact confidence intervals (Clopper & Pearson, 1934) as a function of the true data-generating model $\mathcal{H}_0$ (left panel A) or $\mathcal{H}_1$ (right panel B), the nominal error rates α and β, respectively, and the scale parameter of the Cauchy prior under the alternative hypothesis.

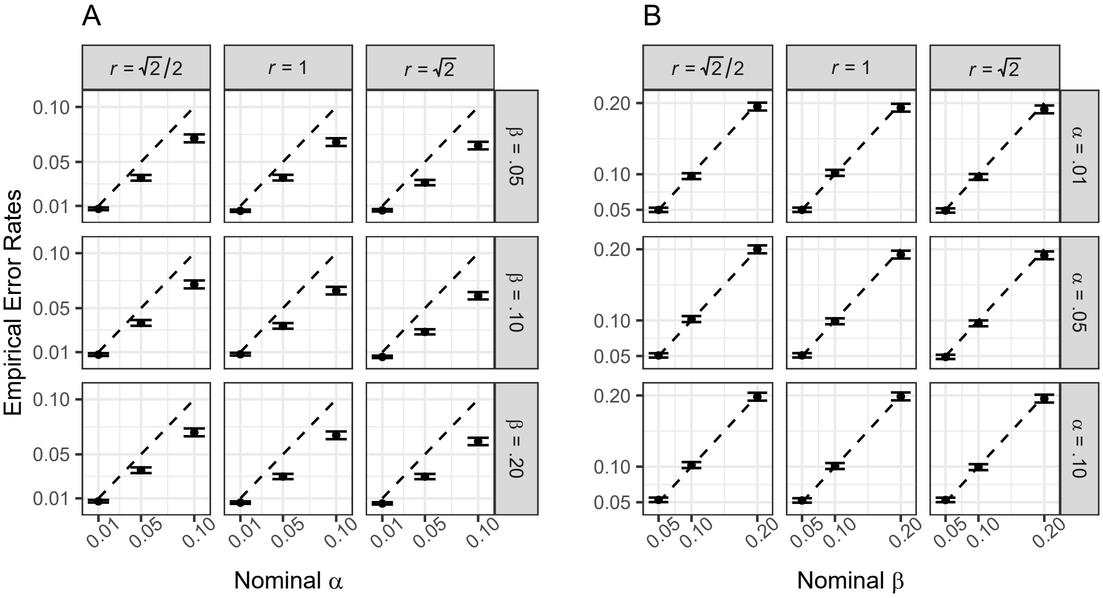

Figure 4. Results of the two-sample Waldian t-test simulation. Empirical error rates are depicted as a function of the nominal error rates and the scale parameter r of the Cauchy prior distribution under the alternative hypothesis. Number of replications per parameter combination: $k = 20,000$. Error bars represent 95% Clopper-Pearson exact confidence intervals (Clopper & Pearson, 1934). **A** The data-generating model corresponds to the null hypothesis: $\delta = 0$. **B** The data-generating model corresponds to the alternative hypothesis: $\delta \sim$ Cauchy(r).

If the null hypothesis is true (panel A), Waldian t tests exert strict control of the probability of a Type-1 error. The observed error rates approximate the nominal error probabilities and never exceed them. In fact, most of the observed rates are significantly lower than the nominal rates, demonstrating the notable influence of overshooting. The deviation, however, is only small. On average, the empirical rates are 34% smaller than the nominal error rates.

If the alternative hypothesis represents the data-generating model (panel B), empirical rates of Type-2 errors almost perfectly match the nominal error rates. Thus, the statistical power of the test (i.e., the complement of the Type-2 error probability, $1 - \beta$) is not negatively affected by its conservative behavior. To summarize, the simulation demonstrates the excellent performance of Waldian t tests in controlling error probabilities of statistical decisions with Bayesian t tests, as derived analytically by Wald (1947).

Reinterpretation of Conventional Criteria

A common way of communicating the evidential strength of a Bayes factor is due to Jeffreys (1961), see also, Lee and Wagenmakers (2013). Often, the continuous scale is divided into discrete categories to provide a heuristic guideline for summarizing and interpreting Bayes factors. According to this taxonomy, a Bayes factor greater than 3 (or, correspondingly, $BF \leq 1/3$) denotes *moderate* evidence, $BF \geq 10$ ($BF \leq 1/10$) represents *strong* evidence, and $BF \geq 30$ ($BF \leq 1/30$) can be interpreted as *very strong* evidence (see Table 1).

This scheme provides an approximate, often-used heuristic to communicate a continuous measure of evidence in categorical terms. However, the specific threshold values of the categories might serve as anchors for researchers aiming at sufficient evidence for a decision about the hypotheses. For example, if one decided to reject or accept the null hypothesis as soon as they observed $BF_{10} \geq 10$ or $BF_{01} \geq 10$, respectively (e.g., Schönbrodt et al., 2017), the verbal label associated with this boundary value might convey misleading impressions as to the quality of this sequential test procedure. For an informed judgment, the long-run properties must be taken into account.

To assess the long-run error rates of a sequential Bayesian t test when adopting the category threshold values suggested by Jeffreys (1961) as decision boundaries, we simply invert the formulae derived by Wald (1947) to find the error probabilities that correspond to these threshold values. Let BF_u denote the upper threshold (leading to acceptance of $\mathcal{H}_1$) and BF_l the lower threshold (leading to acceptance of $\mathcal{H}_0$) of the sequential procedure. According to Wald's formulae, $BF_u = (1 - \beta)/\alpha$ and $BF_l = \beta/(1 - \alpha)$. Solving these two equations for the two unknown error probabilities α^* and β^*, we obtain

$$\alpha^* = \frac{BF_l - 1}{BF_l - BF_u} \tag{15}$$

and

$$\beta^* = \frac{BF_u \cdot BF_l - BF_l}{BF_u - BF_l}. \tag{16}$$

Typically, one would opt for symmetric thresholds in Bayesian statistics, that is, $BF_l = 1/BF_u$, such that the resulting test procedure has symmetric error probabilities. In this case, the formulae presented above reduce to

$$\alpha^* = \beta^* = \frac{1}{BF_u + 1}. \tag{17}$$

According to Equation 17, a sequential Bayesian t test with symmetric thresholds of 10 and 1/10 is associated with error rates $\alpha^* = \beta^* = .09$. These Type-1 and Type-2 error rates might seem surprisingly high when considering that the corresponding verbal label implies "strong evidence" 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).

Clearly, the error rates associated with a sequential Bayesian t test in combination with Bayes factor thresholds of 10 and 1/10, respectively, are considerably larger than those researchers typically aim at. Less surprisingly, employing a threshold denoting "moderate evidence" ($BF_u = 3$) implies even larger error rates: $\alpha^* = \beta^* = .25$ (cf. Schönbrodt et al., 2017). This endorses Schönbrodt et al.'s (2017, p. 332) recommendation to avoid such low threshold values because of their high risk of resulting in incorrect decisions. Table 1 summarizes error rates implied by certain thresholds, as well as threshold values required to satisfy certain error rates.

Table 1
Association of threshold values and error rates for sequential Bayesian t tests

$\mathcal{H}_1$ Threshold		$\mathcal{H}_0$ Threshold			
BF_{10}	Interpretation of Evidence	$1/BF_{10}$	Interpretation of Evidence	α^*	β^*
3.0	moderate	3.00	moderate	0.25	0.25
8.0	moderate	4.50	moderate	0.10	0.20
9.0	moderate	9.00	moderate	0.10	0.10
9.5	moderate	18.00	strong	0.10	0.05
10.0	strong	10.00	strong	0.09	0.09
16.0	strong	4.75	moderate	0.05	0.20
18.0	strong	9.50	moderate	0.05	0.10
19.0	strong	19.00	strong	0.05	0.05
30.0	very strong	30.00	very strong	0.03	0.03
80.0	very strong	4.95	moderate	0.01	0.20
90.0	very strong	9.90	moderate	0.01	0.10
95.0	very strong	19.80	strong	0.01	0.05

Note. BF_{10} = Bayes factor denoting the ratio of the marginal likelihood of $\mathcal{H}_1$ to the marginal likelihood of $\mathcal{H}_0$. Note that $1/BF_{10} = BF_{01}$. α^*, β^* = Type-1 and Type-2 error rates associated with threshold values, respectively. Interpretation of thresholds according to Jeffreys (1961). Note that error rates apply to the sequential procedure and do not consider effects of overshoot.

It is important to note that Equations 15 and 16 define approximate, unconditional error rates of the sequential procedure associated with certain thresholds BF_u and BF_l, not exact error rates conditional on a particular observed result $BF_x > BF_u$ or $BF_x < BF_l$. Like the formulae defining threshold values for the SPRT (Wald, 1947), the formulae presented herein ignore potential overshoot at the point of termination (Berger et al., 1999). Thus, α^* and β^* denote upper bounds to the exact error rates. However, our simulations showed that the overshoot is not crucial and that empirical error rates of the procedure approximate the nominal rates quite well. Consequently, Equations 15 and 16 are useful tools to evaluate the long-run properties of a sequential Bayesian t test with threshold values BF_u and BF_l chosen in accordance with some heuristic taxonomy.

Discussion

Bayesian statistics have become considerably popular among psychologists, fostered by the replication crisis, current advances in computational methods, and persistent critique of classical approaches to statistical inference (Etz & Vandekerckhove, 2018; Jeon & De Boeck, 2017). A particularly influential milestone in this development was the article on Bayesian *t* tests as an alternative to NHST by Rouder et al. (2009). These authors developed and proposed Bayes factors for common *t*-test scenarios. Bayesian *t* tests possess a number of desirable properties and satisfy important desiderata for good statistical procedures. Most importantly, unlike NHST, they allow to measure evidence in favor of the null hypothesis.

Notwithstanding these favorable features, Bayesian *t* tests have a severe limitation: As a continuous measure of evidence, the Bayes factor does not provide a natural basis for binary decisions. In many research contexts, however, such decisions are required, and specific actions are taken depending on whether the null is accepted or rejected. Assuming that we are testing theories about the true state of the world, a single statistical decision can always be right or wrong. There is no way to tell with certainty whether a single decision is wrong. Nevertheless, from the perspective of a cumulative science it is vital that the proportion of decision errors in the long run does not exceed an acceptable limit. This is the very intention of frequentist theories of statistical inference such as the Neyman-Pearson or Wald's theory: to design statistical procedures such that the probabilities of decision errors, conditional on the hypotheses tested being true, can be controlled explicitly (Neyman, 1977; Neyman & Pearson, 1933; Wald, 1947).

There have been previous efforts in the literature to control error probabilities for statistical decisions in the context of Bayesian hypothesis tests (e.g., Berger et al., 1994; Berger et al., 1997; Berger et al., 1999; Gu, Hoijtink, & Mulder, 2016; Schönbrodt & Wagenmakers, 2018). In this article, we presented a sequential extension of Bayesian *t* tests based on Wald's sequential probability ratio test. These Waldian *t* tests (a) are based on an analytically derived framework of error probability control, (b) are easily applied to any Bayesian *t* test based on proper priors, and (c) allow for efficient hypothesis testing with strict control of error rates. Importantly, Waldian *t* tests also maintain a fully Bayesian justification. The interpretation of the numerical value of the Bayes factor obtained after reaching one of the a priori defined thresholds is not affected by the sequential application. Therefore, we believe that Waldian *t* tests remedy a notable limitation of Bayesian *t* tests by combining their beneficial properties with those of classical statistical procedures.

Limitations

Waldian *t* tests are based on the assumption that sampling is continued until a decision threshold is reached, upon which the procedure is terminated and one of the two hypotheses is accepted. The assumption that sampling can be continued indefinitely is vital in any sequential procedure designed for controlling error probabilities. 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, 2019; but see Armitage, 1957). 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 application 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 (e.g., Cohen, 1988) or a closed sequential test that maintains a pre-specified maximum sample size (e.g., Proschan, Lan, & Wittes, 2006) might be appropriate alternatives. However, these approaches require that both the null and the alternative hypothesis can be specified as simple point hypotheses. If researchers feel more comfortable with a composite alternative hypothesis as employed in Bayesian and Waldian t tests, a simulation-based solution might be more appropriate to determine the necessary (fixed) sample size and critical value for a Bayesian t test such that the procedure satisfies prespecified error rates. Such an approach, referred to as *Bayes factor design analysis* (BFDA), has been developed and introduced by Schönbrodt and Wagenmakers (2018). Although BFDA is rooted in the Bayesian framework and focuses on strength of evidence rather than error rates, it can be used to assess any kind of long-run property of the test procedure by means of simulation, assuming a certain true state of the world. Therefore, if an open sequential procedure is not an option, BFDA can be used to design Bayesian t tests such that they allow for error control (see Stefan, Gronau, Schönbrodt, & Wagenmakers, 2019, for a tutorial on BFDA).

Furthermore, it is important to keep in mind that the error probabilities in a Waldian t test denote the weighted average probability of decision errors across all possible parameter values defined by the alternative hypothesis. This is in line with the requirements specified in Equation 7 and makes sense, for example, if we assume the composite hypothesis to represent a range of possible, differently weighted effect sizes (e.g., in case of a theoretically motivated null hypothesis). However, the focus on weighted average probabilities also implies that the specified error probability for the alternative hypothesis is not constant for all parameter values $\delta \in \Delta_1$. If this is required, one would need to find a suitable weight function (i.e., prior distribution), for which the resulting test would satisfy this requirement. However, the construction of such a weight function on δ can be complex (Wald, 1947) and, with respect to the Bayes factor, would possibly result in a prior distribution without a substantive interpretation (Gu et al., 2016).

A different approach would be to define a minimum value $\delta_{\text{min}} \in \Delta_1$ for which the requirement can be imposed that the sequential test has error probability less or equal to β for any δ greater than or equal to δ_{min}. In this case, an SPRT t test such as Hajnal's t test (Schnuerch & Erdfelder, 2019) or a Neyman-Pearson t test based on a point alternative hypothesis $\mathcal{H}_1$: $\delta = \delta_{\text{min}}$ will satisfy the error requirement, that is, $P(\text{reject } \mathcal{H}_1|\delta) \leq \beta$, $\forall \delta \in \{\Delta_1 | \delta \geq \delta_{\text{min}}\}$.

This characteristic of both Hajnal's t test and Neyman-Pearson t tests is highly desirable because it enables researchers to test a composite hypothesis (i.e., $\delta \geq \delta_{\text{min}}$) with a specific upper-bound error probability. However, the appealing evidential interpretation of the specified likelihood ratio (as well as the Bayes factor) is no longer valid in this case. The ratio always denotes the relative evidence for the specified models. If the alternative model is based on δ_{min}, the likelihood ratio denotes the relative evidence for the hypothesis that $\delta = \delta_{\text{min}}$, not that $\delta \geq \delta_{\text{min}}$. In other words, the likelihood ratio (as well as the Bayes factor) can be interpreted as relative evidence only for the specified statistical models, but no longer for the underlying psychological hypotheses.

Conclusion

From a Bayesian perspective, prior distributions represent uncertainty or subjective belief about more or less plausible values of the parameters. The Bayes factor, in turn, denotes how the data inform and change subjective belief about different hypotheses. In the classical frequentist framework, in contrast, a prior distribution has its counterpart in a random effect, that is, variation in true parameter values or effect sizes across experiments. Such an assumption is reasonable, for example, when a theoretically motivated simple null hypothesis ($\delta = 0$) is tested against a composite, not further specified alternative hypothesis ($\delta \neq 0$). With Waldian *t* tests, these hypotheses can be tested with reliable error probability control in a classical (Neyman-Pearson-Wald) sense.

It is not our intention to suggest that psychologists should abandon hypothesis tests based on simple hypotheses. Eventually, the choice of a statistical model should be determined by the psychological hypotheses and the aim of the statistical test. If the theory at test predicts a specific (minimum) effect size or if decision error probability control is required for effect sizes equal to or more extreme than some minimum relevant effect size (e.g., in various types of psychological intervention research), this warrants the specification of point hypotheses and the use of Neyman-Pearson *t* tests or (notably more efficient) sequential tests such as Hajnal's *t* test (Schnuerch & Erdfelder, 2019).

If, however, the psychological theory predicts an invariance, which is typically represented by a point null hypothesis ($\delta = 0$), any deviation from this point (i.e., any $\delta \neq 0$) would contradict the theory (Rouder et al., 2009). In this case, the Cauchy prior distribution has a reasonable substantive interpretation as a weight function for plausible, non-zero effect sizes under the alternative hypothesis. Thus, for the case of a substantively motivated null hypothesis tested against an unrestricted and not further specified alternative hypothesis, Bayesian *t* tests with default priors as suggested by Rouder et al. (2009) might be more appropriate than tests based on a point alternative hypothesis. When used as a means to accept or reject the null hypothesis with controlled error probabilities, however, Waldian *t* tests are a useful and easy-to-apply extension of Bayesian *t* tests.

Statistical inference aiming at error probability control (i.e., classical statistics) on the one hand and updating of subjective beliefs (i.e., Bayesian statistics) on the other hand are rooted in two fundamentally different approaches to probability and statistics. They might often be perceived as incompatible, although they are, in fact, "both quite legitimate" (Efron, 2005, p. 1). Therefore, with this article, we hope to bring together beneficial properties from both worlds and combine them in the Waldian *t* tests.

Open Practices Statement

The simulation data reported in this manuscript, as well as all R scripts to reproduce the simulation and analysis are available at the Open Science Framework (https://osf.io/z5vsy/?view_only=63252a8bd2374b8297bf8bb9fdd12442).

References

Armitage, P. (1957). Restricted sequential procedures. *Biometrika*, *44*, 9–26. doi:10.2307/2333237

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*, 235–244. doi:10.2307/2343787

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

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. doi:10.1016/j.jmp.2015.12.007

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

Berger, J. O., & Bayarri, M. J. (2004). The interplay of Bayesian and frequentist analysis. *Statistical Science*, *19*, 58–80. doi: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. doi:10.1214/ss/1030037904

Berger, J. O., Boukai, B., & Wang, Y. (1999). Simultaneous Bayesian-frequentist sequential testing of nested hypotheses. *Biometrika*, *86*, 79–92. doi: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*, 1787–1807. doi:10.1214/aos/1176325757

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

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*, 1–38. doi:10.5334/joc.72

Clopper, C. J., & Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, *26*, 404–413. doi:10.2307/2331986

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

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

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

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

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

Efron, B. (2005). Bayesians, frequentists, and scientists. *Journal of the American Statistical Association*, *100*, 1–5. doi:10.1198/016214505000000033

Erdfelder, E., & Heck, D. W. (in press). Detecting evidential value and p-hacking with the p-curve tool: A word of caution. *Zeitschrift für Psychologie/Journal of Psychology.*

Etz, A., & Vandekerckhove, J. (2018). Introduction to Bayesian inference for psychology. *Psychonomic Bulletin & Review*, *25*, 5–34. doi:10.3758/s13423-017-1262-3

Etz, A., & Wagenmakers, E.-J. (2017). J. B. S. Haldane's contribution to the Bayes factor hypothesis test. *Statistical Science*, *32*, 313–329. doi:10.1214/16-STS599

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.

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*, 587–606. doi:10.1016/j.socec.2004.09.033

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

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

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

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

Kiefer, J. (1977). Conditional confidence statements and confidence estimators. *Journal of the American Statistical Association*, *72*, 789–808. doi:10.1080/01621459.1977.10479956

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*, 187–192. doi: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. doi:10.1016/j.jmp.2015.06.004

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. doi:10.1016/j.jmp.2015.11.001

Morey, R. D., & Rouder, J. N. (2015). BayesFactor: Computation of Bayes factors for common designs. Retrieved from https://CRAN.R-project.org/package=BayesFactor

Morey, R. D., Rouder, J. N., Verhagen, J., & Wagenmakers, E.-J. (2014). Why hypothesis tests are essential for psychological science: A comment on Cumming (2014). *Psychological Science*, *25*, 1289–1290. doi:10.1177/0956797614525969

Neyman, J. (1977). Frequentist probability and frequentist statistics. *Synthese*, *36*, 97–131. doi: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*, 289–337. doi:10.1098/rsta.1933.0009

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

R Core Team. (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from http://www.R-project.org/

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

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

Rouder, J. N., Morey, R. D., Verhagen, J., Province, J. M., & Wagenmakers, E.-J. (2016). Is there a free lunch in inference? *Topics in Cognitive Science*, *8*, 520–547. doi:10.1111/tops.12214

Rouder, J. N., Morey, R. D., & Wagenmakers, E.-J. (2016). The interplay between subjectivity, statistical practice, and psychological science. *Collabra*, *2*, 1–12. doi: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. doi:10.3758/PBR.16.2.225

Royall, R. (1997). *Statistical evidence: A likelihood paradigm.* Abingdon: Routledge.

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

Sanborn, A. N., & Hills, T. T. (2014). The frequentist implications of optional stopping on Bayesian hypothesis tests. *Psychonomic Bulletin & Review*, *21*, 283–300. doi: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*, 309–311. doi:10.3758/s13423-014-0607-4

Schnuerch, M., & Erdfelder, E. (2019). Controlling decision errors with minimal costs: The sequential probability ratio t test. *Psychological Methods.* Advance online publication. doi:10.1037/met0000234

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, *25*, 128–142. doi:10.3758/s13423-017-1230-y

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*, *22*, 322–339. doi:10.1037/met0000061

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*, 534–547. doi:10.1037/a0033242

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

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

Wagenmakers, E.-J., & Gronau, Q. F. (2018). *Error rate schmerror rate.* Bayesian Spectacles. Retrieved from https://www.bayesianspectacles.org/error-rate-schmerror-rate/

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., … Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*, 35–57. doi:10.3758/s13423-017-1343-3

Wagenmakers, E.-J., Verhagen, J., Ly, A., Matzke, D., Steingroever, H., Rouder, J. N., & Morey, R. D. (2017). The need for Bayesian hypothesis testing in psychological science. In S. O. Lilienfeld & I. D. Waldman (Eds.), *Psychological science under scrutiny* (pp. 123–138). Hoboken, NJ: Wiley.

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

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

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

Wetherill, G. B. (1975). *Sequential methods in statistics* (2. ed.). London: Chapman and Hall.

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

Zellner, A., & Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. *Trabajos de Estadistica Y de Investigacion Operativa*, *31*, 585–603. doi:10.1007/BF02888369

Sequential Hypothesis Tests for Multinomial Processing Tree Models

Martin Schnuerch and Edgar Erdfelder
University of Mannheim

Daniel W. Heck
Philipps-Universität Marburg

Stimulated by William H. Batchelder's seminal contributions in the 1980s and 1990s, multinomial processing tree (MPT) modeling has become a powerful and frequently used method in various research fields, most prominently in cognitive psychology and social cognition research. MPT models allow for estimation of, and statistical tests on, parameters that represent psychological processes underlying responses to cognitive tasks. Therefore, their use has also been proposed repeatedly for purposes of psychological assessment, for example, in clinical settings to identify specific cognitive deficits in individuals. However, a considerable drawback of individual MPT analyses emerges from the limited number of data points per individual, resulting in estimation bias, large standard errors, and low power of statistical tests. Classical test procedures such as Neyman-Pearson tests often require very large sample sizes to ensure sufficiently low Type 1 and Type 2 error probabilities. Herein, we propose sequential probability ratio tests (SPRTs) as an efficient alternative. Unlike Neyman-Pearson tests, sequential tests continuously monitor the data and terminate when a predefined criterion is met. As a consequence, SPRTs typically require only about half of the Neyman-Pearson sample size without compromising error probability control. We illustrate the SPRT approach to statistical inference for simple hypotheses in single-parameter MPT models. Moreover, a large-sample approximation, based on ML theory, is presented for typical MPT models with more than one unknown parameter. We evaluate the properties of the proposed test procedures by means of simulations. Finally, we discuss benefits and limitations of sequential MPT analysis.

Keywords: multinomial processing tree models, hypothesis tests, efficiency, sequential analysis, sequential probability ratio test

1 Multinomial Processing Tree Models

Among a multitude of outstanding contributions to the field of psychology, one of the arguably most prominent instances of William H. Batchelder's (1940–2018) scientific impact is the development of a class of stochastic models for the measurement of cognitive processes, known as multinomial processing tree (MPT) models. In what is now considered a classical article, Riefer and Batchelder (1988) introduced and promoted the use of MPT models which, in contrast to other scientific areas, had received but little attention in psychology at the time (Erdfelder et al., 2009). Stimulated by this pioneering work and Batchelder's ongoing effort in the following years (e.g., Batchelder & Riefer, 1999), MPT models have become a powerful instrument to measure and disentangle the contribution of latent processes underlying observed behavior.

MPT models are substantively motivated stochastic models for categorical data (but see Heck, Erdfelder, & Kieslich, 2018, for an extension to continuous data). They are based on the assumption that each observable response in a specific paradigm originates from a finite set of sequences of discrete processing states. These sequences are conceptualized as branches in a processing tree. Nodes along these branches denote latent cognitive states and the links between the nodes represent (conditional) probabilities of entering the respective states. The product of these link probabilities determines the branch probability. Each category probability, in turn, is defined as the sum of probabilities of all branches terminating in this category. Based on the assumption that observed category frequencies follow a multinomial distribution, the (conditional) link probabilities can be estimated and, thus, the contribution of each latent processing state can be measured and tested statistically (Erdfelder et al., 2009; Hu & Batchelder, 1994).

Nowadays, MPT models are widely used in various branches of psychology, particularly in (social-)cognitive research. Even though the primary context of MPT applications is experimental psychology, Batchelder himself repeatedly promoted the use of MPT models for psychometric purposes (e.g., Batchelder, 1998; Batchelder & Riefer, 1999). Unlike item response models, for example, MPT models are based on explicit assumptions about the latent cognitive processes underlying observed responses and aim at measuring and disentangling these processes. Thus, Batchelder (1998) identified an "untapped potential" (p. 331) of what he referred to as "cognitive psychometrics" for individual assessments of specific cognitive processes, for example in clinical settings.

Despite the apparent appeal of MPT models for individual assessment, there is a notable limitation of this type of cognitive psychometrics. In experimental settings, MPT analyses typically make use of group data, either in a pooled or a hierarchical fashion (Chechile, 2009; Heck, Arnold, & Arnold, 2018; Klauer, 2006, 2010; Smith & Batchelder, 2010). As a consequence, parameter estimates and statistical tests are based on many data points, often resulting in high precision of estimates and high statistical power of tests. Individual analyses, in contrast, are typically based on far fewer observations. Thus, parameter estimates may be biased and will necessarily be less precise, resulting in large standard errors and low statistical power (Batchelder, 1998).

To remedy the problem of few observations in individual parameter estimation, Batchelder suggested to make use of Bayesian methods. Drawing on data from other individuals that are matched to the testee based on theoretical considerations (e.g., a reference group similar in age and educational background), one can construct a prior distribution for the parameters of interest. Using Bayes' theorem, this prior is then combined with the testee's

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

This research was supported by a grant from the Deutsche Forschungsgemeinschaft (DFG, GRK 2277) to the Research Training Group "Statistical Modeling in Psychology" (SMiP).

Reproducible scripts for simulations and analysis as well as all simulated raw data can be downloaded from the Open Science Framework (https://osf.io/98erb/).

Correspondence concerning this article should be addressed to Martin Schnuerch or Edgar Erdfelder, Cognition and Individual Differences Lab, Department of Psychology, University of Mannheim, 68131 Mannheim, Germany. E-mail: martin.schnuerch@psychologie.uni-mannheim.de or erdfelder@psychologie.uni-mannheim.de

data to obtain the posterior distribution. The mean (or mode) of this distribution serves as a point estimate while its variance or other measures of dispersion denote estimation uncertainty. When there is little variance in the prior, this empirical Bayes estimation procedure will result in a more precise estimate than classical maximum likelihood estimation without prior information (Batchelder, 1998). Of course, there is also a danger of considerable bias if the testee deviates systematically from other individuals (i.e., if the prior is misspecified).

A frequent goal in individual clinical assessment, however, goes beyond mere estimation of model parameters: To identify a specific cognitive deficit or to decide on a particular intervention for the individual testee, statistical tests on model parameters are required. For example, to assess whether or not an individual is able to utilize a certain cognitive process, one might want to test whether the corresponding parameter is substantially different from zero. In MPT modeling, tests of parameter constraints typically rely on null-hypothesis significance testing (NHST) based on the asymptotic distribution of some fit statistic under the null hypothesis (Batchelder & Riefer, 1999). In MPT models, these fit statistics denote the distance between model-implied and observed category frequencies. They can be characterized as a power divergence family (Read & Cressie, 1988), the most well-known special cases of which are Pearson's χ^2 or the log likelihood ratio G^2 (Hu, 1999; Hu & Phillips, 1999).

Standard applications of NHST to MPT models typically ignore statistical power, that is, the probability of rejecting a set of parameter constraints if the constraints do indeed not hold in the population. However, both in basic research and in clinical settings, sufficient statistical power is necessary for unbiased inference (Batchelder & Riefer, 1990, 1999). To this end, classical methods to control statistical error probabilities based on the seminal theory by Neyman and Pearson (1933) require an a priori power analysis. Given a certain expected effect size and a predefined Type 1 error probability α, the Type 2 error probability β (and the power of the test, $1-\beta$) is a function of the sample size. Although power analyses are easily carried out with MPT software (e.g., multiTree; Moshagen, 2010) or general-purpose software for power analysis (e.g., G*Power; Faul, Erdfelder, Buchner, & Lang, 2009), there are two major drawbacks in the context of MPT analyses: First, a power analysis not only requires assumptions concerning the test-relevant parameters as specified by the null and the alternative hypothesis but depends on all other model parameters as well. This poses a problem whenever the model contains parameters for which the population values are unknown, so-called *nuisance* parameters. The second major limitation of classical power analyses in the Neyman-Pearson framework concerns scenarios in which the expected effect size is small. In this case, classical Neyman-Pearson tests require extremely large numbers of observations, often much larger than realistically feasible.

The problem of achieving a sufficiently powered hypothesis test is particularly pressing when data collection is costly: either when assessing a single participant with as few trials as possible or when each participant provides only a single data point (e.g., Batchelder, 1998; Heck, Thielmann, Moshagen, & Hilbig, 2018; Klauer, Stahl, & Erdfelder, 2007; Moshagen, Hilbig, Erdfelder, & Moritz, 2014; Moshagen, Musch, & Erdfelder, 2012; Schild, Heck, Ścigała, & Zettler, 2019). However, it potentially applies to any MPT model analysis (Batchelder & Riefer, 1990, 1999; Riefer & Batchelder, 1988). Therefore, in this article we introduce a sequential statistical method for hypothesis testing in MPT models that (1) allows to control both α and β error probabilities (unlike NHST), (2) requires on average much less observations than classical power analyses, and (3) does not rest on explicit assumptions about the population values of nuisance parameters of the model.

The approach we promote herein is based on Abraham Wald's sequential probability ratio test

(Wald, 1947). In the following, we introduce the basic idea as well as an extension of Wald's method by D. R. Cox (1963). We then show how sequential tests can be used for efficient hypothesis tests in MPT models and how this may improve the applicability of MPT models for purposes of individual assessment. Overall, with the present article we hope to increase efficiency not only of typical experimental applications of MPT models, but also for applications to individuals in the context of cognitive psychometrics.

2 Sequential Analysis

2.1 Sequential Probability Ratio Tests

Classical statistical methods rely on fixed samples of an a priori defined size. Sequential statistics, in contrast, are based on the continuous monitoring of the data throughout the sampling process. This process continues until some predefined criterion is met, at which point sampling is terminated (*optional stopping*) and a statistical decision is made. Crucially, unlike the recursive application strategy of classical methods known as p-hacking (Simmons, Nelson, & Simonsohn, 2011), sequential methods do not compromise control of long-term error rates (Wetherill, 1975).

Due to their characteristic to terminate early whenever the data strongly support a hypothesis, statistical analysis may substantially reduce the required sample size. For a decision between two simple hypotheses, Wald's (1947) sequential probability ratio test (SPRT) has been proven to be the most efficient test (Matthes, 1963; Wald & Wolfowitz, 1948). That is, for given long-term error rates α and β, there is no test procedure that requires less observations than the SPRT on average.

To illustrate the SPRT, consider a random variable $\mathbf{X}$, $X \sim f(x|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ denotes the true parameter vector of the underlying population. The random variable may be discrete or continuous, in which case the function $f(.)$ refers to the probability mass or the probability density, respectively. Assume a test of the two simple hypotheses $\mathcal{H}_0$: $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ versus $\mathcal{H}_1$: $\boldsymbol{\theta} = \boldsymbol{\theta}_1$. A hypothesis is *simple* when all parameters of the statistical model that define the probability distribution of the data are either known or specified by the hypothesis. If at least one parameter is not known or restricted to a specific value, the hypothesis is *composite*. For the given example, the parameter vector $\boldsymbol{\theta}$ is completely specified under each hypothesis. Thus, the hypotheses are simple and the SPRT is the optimal test to decide between them with a given strength (α, β).

In the SPRT, the ratio of the probabilities of the observed data after any n^{th} observation, $x^n = (x_1, ..., x_n)$, under each hypothesis i is computed. As the probability density is proportional to the likelihood, that is, $f(x^n|\boldsymbol{\theta}_i) \propto \mathcal{L}(\boldsymbol{\theta}_i;\ x^n)$, this ratio is typically referred to as a likelihood ratio (LR):

$$LR_n = \frac{f(x^n|\boldsymbol{\theta}_1)}{f(x^n|\boldsymbol{\theta}_0)} = \frac{\mathcal{L}(\boldsymbol{\theta}_1;\ x^n)}{\mathcal{L}(\boldsymbol{\theta}_0;\ x^n)}. \tag{1}$$

Sampling continues by adding independent observations x_{n+1} as long as

$$B < LR_n < A. \tag{2}$$

If $LR_n \geq A$, sampling is terminated and $\mathcal{H}_1$ is accepted. By definition, any sample x^n which leads to the acceptance of $\mathcal{H}_1$ is thus at least A times as likely under $\mathcal{H}_1$ as under $\mathcal{H}_0$. This implies that the probability to accept $\mathcal{H}_1$ is at least A times larger under $\mathcal{H}_1$ than under $\mathcal{H}_0$. In the usual notation based on the Neyman-Pearson theory, the former probability is defined as $1 - \beta$, whereas the latter is denoted by α. Hence, $1 - \beta \geq A\alpha$, which can be written as

$$\frac{1-\beta}{\alpha} \geq A. \tag{3}$$

In contrast, if $LR_n \leq B$, sampling is terminated and $\mathcal{H}_0$ is accepted. Following the same logic as for A, we see that

$$\frac{\beta}{1-\alpha} \leq B, \tag{4}$$

which implies that upper/lower limits for A and B are given by $(1-\beta)/\alpha$ and $\beta/(1-\alpha)$, respectively.

In practical applications, however, the inequalities in (3) and (4) can be treated as equalities defining threshold values A and B for the LR that satisfy pre-specified statistical error probabilities α and β. More precisely, the resulting sequential test procedure provides an approximate control of error probabilities with α and β serving as upper bounds to the actual error rates (Wald, 1947; Wetherill, 1975).

Functions describing the test procedure's properties (i.e., power and expected sample size at termination) can be approximated analytically by formulae derived by Wald (1947). Moreover, as mentioned before, the SPRT has been proven to be the most efficient test for given error rates. As soon as the statistical model defining the probability distribution of the data contains nuisance parameters, however, the general theory of the SPRT no longer applies.

This constitutes a practically relevant limitation since composite hypotheses due to nuisance parameters occur frequently in many common MPT models (e.g., in models for memory paradigms, which typically comprise guessing parameters). Different methods have been proposed to remedy this problem: For example, Wald (1947) suggested to integrate out nuisance parameters by means of weight functions (that resemble prior distributions in Bayesian inference). In a different approach, the likelihood ratio is constructed based on simple sufficient statistics (Barnard, 1952; D. R. Cox, 1952; Rushton, 1950). Although this approach provides an adequate solution for certain problems such as the classical *t* test (Schnuerch & Erdfelder, 2019), its applicability is restricted to specific situations. In the following, we consider a more general method introduced by D. R. Cox (1963), building on Bartlett's (1946) idea to construct a sequential test based on asymptotic maximum likelihood (ML) theory.

2.2 Sequential Maximum Likelihood Ratio Tests

Let **X** be a random variable denoting the observed data, with $X \sim f(x|\boldsymbol{\theta}, \boldsymbol{\phi})$. Similar as in the SPRT above, we consider a test of the hypotheses $\mathcal{H}_i$: $\boldsymbol{\theta} = \boldsymbol{\theta}_i$ ($i = 0, 1$), $\boldsymbol{\phi}$ denoting nuisance parameters of the statistical model. The method developed by D. R. Cox (1963) and outlined in this section applies to any $\boldsymbol{\theta}$, $\boldsymbol{\phi}$ regardless of their dimensionalities. Therefore, without loss of generality, we will assume that both parameters are single-valued in what follows. A detailed mathematical justification of D. R. Cox's method can be found in Breslow (1969).

In the SPRT, it is straightforward to consider the log likelihood ratio rather than the likelihood ratio. Assume the true value of ϕ was known, then the SPRT as defined in the previous section would require to continue sampling as long as

$$\log\left(\frac{\beta}{1-\alpha}\right) < \ell(\theta_1, \phi;\ x^n) - \ell(\theta_0, \phi;\ x^n) < \log\left(\frac{1-\beta}{\alpha}\right), \tag{5}$$

where $\ell(\theta_i, \phi;\ x^n)$ denotes the log likelihood for hypothesis i after n observations. Calculations involving exact log-likelihood functions are often difficult or even infeasible. As a remedy, based on large-sample theory, the exact log likelihood can be replaced by a second-order Taylor series expansion about the true parameter value θ, treating the difference $\theta_i - \theta$ ($i = 0, 1$) as of order $1/\sqrt{n}$ (cf. Joanes, 1972):

$$\begin{aligned}\ell(\theta_i, \phi;\ x^n) = \ell(\theta, \phi;\ x^n) + (\theta_i - \theta)\frac{\partial \ell(\theta, \phi;\ x^n)}{\partial \theta} \\ + \tfrac{1}{2}(\theta_i - \theta)^2 \frac{\partial^2 \ell(\theta, \phi;\ x^n)}{\partial \theta^2}\ ,\end{aligned} \tag{6}$$

such that the log likelihood ratio in (5) becomes

$$\begin{aligned}&(\theta_1 - \theta_0)\frac{\partial \ell(\theta, \phi;\ x^n)}{\partial \theta} \\ &+ \tfrac{1}{2}(\theta_1 - \theta_0)(\theta_1 + \theta_0 - 2\theta)\frac{\partial^2 \ell(\theta, \phi;\ x^n)}{\partial \theta^2}.\end{aligned} \tag{7}$$

If, in contrast, ϕ is not known, the log likelihood ratio can be constructed using the ML estimate $\hat{\phi}$ based on x^n, that is,

$$\ell(\theta_1, \hat{\phi};\ x^n) - \ell(\theta_0, \hat{\phi};\ x^n)\,. \tag{8}$$

Note that Bartlett (1946) suggested separate ML estimates for the nuisance parameter ϕ conditional on $\mathcal{H}_1$ and $\mathcal{H}_0$ (i.e., the estimates $\hat{\phi}_1$ and $\hat{\phi}_0$ assuming $\theta = \theta_1$ or $\theta = \theta_0$, respectively). In contrast, D. R. Cox's (1963) method involves the use of a single estimate $\hat{\phi}$ for both terms in (8), conditional on a model without restrictions on θ or ϕ. Expanding about the true parameter values (θ, ϕ) analogously to (6), (8) becomes

$$\begin{aligned}&(\theta_1 - \theta_0)\frac{\partial \ell(\theta, \phi;\ x^n)}{\partial \theta}\\ &+ \tfrac{1}{2}(\theta_1 - \theta_0)(\theta_1 + \theta_0 - 2\theta)\frac{\partial^2 \ell(\theta, \phi;\ x^n)}{\partial \theta^2}\\ &+ (\theta_1 - \theta_0)(\hat{\phi} - \phi)\frac{\partial^2 \ell(\theta, \phi;\ x^n)}{\partial \theta \partial \phi}.\end{aligned} \tag{9}$$

It is easy to see that (9) is equivalent to (7) if the last term becomes 0, that is, if θ and ϕ are independent and, thus,

$$E\left[\frac{1}{n}\frac{\partial^2 \ell(\theta, \phi;\ x^n)}{\partial \theta \partial \phi}\right] \xrightarrow[n\to\infty]{} 0\,. \tag{10}$$

In this case, the ML estimates $\hat{\theta}$ and $\hat{\phi}$ are asymptotically independent as well. A simple SPRT as defined in (5) where ϕ is replaced by $\hat{\phi}$ is then asymptotically equivalent to that when ϕ is known. If $\hat{\theta}$ and $\hat{\phi}$ are not asymptotically independent, however, the test procedure will not satisfy the long-run error rates implied by α and β. As a remedy, the sampling error of $\hat{\phi}$ must be taken into account.

Equation (9) can be further simplified based on large-sample ML theory, showing that it is asymptotically equivalent to the following expression (see Appendix A and D. R. Cox, 1963, for details):

$$n\mathcal{I}_{\theta\theta}(\theta_1 - \theta_0)\left[\hat{\theta} - \tfrac{1}{2}(\theta_0 + \theta_1)\right], \tag{11}$$

where $\mathcal{I}_{\theta\theta}$ denotes the (θ, θ) element or submatrix of the expected Fisher information matrix $\mathcal{I}(\theta, \phi)$ for sample size $n = 1$, assuming observations to be independent and identically distributed.

For simplification, D. R. Cox (1963) suggested to base the sequential test procedure on a monotonic transformation of (11) obtained by dropping the multiplicative constant $\mathcal{I}_{\theta\theta}(\theta_1 - \theta_0)$ (see also Wetherill, 1975, p. 60),

$$T_n = n\left[\hat{\theta} - \tfrac{1}{2}(\theta_0 + \theta_1)\right], \tag{12}$$

where $\hat{\theta}$ is the ML estimate of θ based on x^n. This test statistic has to be computed after any n^{th} observation, and stopping boundaries corresponding to the constant likelihood-ratio boundaries of the SPRT (Equation 2) are given by

$$\frac{\mathcal{V}_{\theta\theta}}{\theta_1 - \theta_0}\log\left(\frac{\beta}{1-\alpha}\right) < T_n < \frac{\mathcal{V}_{\theta\theta}}{\theta_1 - \theta_0}\log\left(\frac{1-\beta}{\alpha}\right). \tag{13}$$

In (13), $\mathcal{V}_{\theta\theta}$ denotes the (θ, θ) element of the inverse of the expected unit Fisher information, that is, $\mathcal{V} = \mathcal{I}(\theta, \phi)^{-1}$. In many cases, the analytical derivation of the expected Fisher information is infeasible. Thus, for practical purposes, it can be replaced by the observed Fisher information $\mathbf{I}(\hat{\theta}, \hat{\phi})$, that is,

$$\mathbf{I}(\hat{\theta}, \hat{\phi}) = -\frac{1}{n}\mathbf{H}(\hat{\theta},\ \hat{\phi})\,, \tag{14}$$

where $\mathbf{H}(\hat{\theta},\ \hat{\phi})$ is the Hessian matrix of second-order partial derivatives of the log likelihood function, evaluated at the ML estimates.

As an element of the inverse of the unit Fisher information, $\mathcal{V}_{\theta\theta}$ (or, when using the observed information matrix, $\mathbf{V}_{\theta\theta}$) denotes the variance of the ML estimate $\hat{\theta}$ based on a single observation (cf. Ly, Marsman, Verhagen, Grasman, & Wagenmakers, 2017). Thus, the threshold values in (13) are adjusted based on the precision with which the test-relevant parameter is estimated, thereby correcting for the uncertainty that results from the necessity to estimate the unknown nuisance parameter ϕ.

D. R. Cox's test, henceforth referred to as sequential maximum likelihood ratio test (SMLRT), satisfies the long-run error rates α and β for testing hypotheses about θ without making explicit assumptions about the nuisance parameters ϕ (D. R. Cox, 1963; Wetherill, 1975). As it is based on asymptotic ML theory, however, the approximations involved in the derivation of the formulae cannot be expected to work sufficiently well for small samples. Hence, the proposed sequential test requires a sufficiently large initial sample (C. P. Cox & Roseberry, 1966). Otherwise, it is not ensured that the Taylor series expansion in (9) is valid or that the observed Fisher information $\mathbf{I}(\hat{\theta}, \hat{\phi})$ provides a good approximation of the expected Fisher information $\mathcal{I}(\theta, \phi)$ (Hu & Phillips, 1999). Nevertheless, even though the initial sample size needs to be sufficiently large, we show below that the SMLRT still requires on average much smaller sample sizes than Neyman-Pearson tests without compromising error probability control.

The practical implementation of the SMLRT for MPT models is straightforward with MPT software such as, for example, multiTree (Moshagen, 2010) or MPTinR (Singmann & Kellen, 2013). After any n^{th} observation, the ML estimate $\hat{\theta}$ can easily be computed with these software packages. Additionally, $\mathbf{V}_{\theta\theta}$ can be computed from software output based on the estimated standard error of $\hat{\theta}$, $SE_{\hat{\theta}}$. Since $SE_{\hat{\theta}}$ is the (θ, θ) element of $\left[n\mathbf{I}(\hat{\theta}, \hat{\phi})\right]^{-1/2}$, it follows that $\mathbf{V}_{\theta\theta} = n(SE_{\hat{\theta}})^2$.

3 Sequential MPT Analysis

3.1 Case 1: Simple Hypothesis

As a running example, consider a psychometric experiment administered to an individual participant in a clinical assessment situation. Assume we are interested in the individual's perceptual abilities. Specifically, we want to assess whether or not the participant is able to detect a visual stimulus of a given intensity.

The experiment is carried out as follows: In the style of classical experiments on visual thresholds (Blackwell, Pritchard, & Ohmart, 1954) and decision processes underlying visual perception (Swets, Tanner, & Birdsall, 1961), the participant is presented with a visual stimulus in one of two defined temporal intervals in each trial. A stimulus typically used in such experiments is a flash of light displayed on a screen (for 100 ms, say) with a certain diameter and magnitude (that is, luminous intensity). Following each trial, the participant is prompted to decide in which of the two intervals the stimulus was presented. Thus, the perceptual performance is measured in a two-alternative forced-choice test (2AFC).

If the participant detects the stimulus, they will answer correctly. If they do not detect the stimulus, however, they might still give a correct answer by guessing the interval in which the stimulus was presented. Thus, the performance in the 2AFC is diluted by guessing processes which do not represent actual perceptual abilities (Swets et al., 1961). In order to assess these directly, the processes underlying response behavior in the 2AFC can be disentangled by means of an MPT model.

Figure 1 displays the simplest instance of an MPT model for the paradigm under consideration. In each trial, participants either enter a state of detection (with probability d) and choose the correct answer, or they do not detect the stimulus ($1-d$). In this state of uncertainty, they have to guess which of the intervals contained the stimulus. Thus, they can either guess correctly (with conditional probability g) or incorrectly ($1 - g$).

Formally, the probability of each branch j (j = 1, ..., J) leading to response category k (k = 1, ..., K) in a binary MPT model is defined as

$$p_{jk}(\mathbf{\Theta}) = c_{jk} \prod_{s=1}^{S} \theta_s^{a_{jks}} (1 - \theta_s)^{b_{jks}} , \qquad (15)$$

where $\mathbf{\Theta} = (\theta_1, ..., \theta_S)$ represents the vector of parameters in the model denoting the (conditional) link probabilities along the branches, with $\mathbf{\Theta} \in \mathbf{\Omega} = [0,\ 1]^S$. The count variables a_{jks} and b_{jks} indicate how often a parameter θ_s (or its complement $1-\theta_s$, respectively) occurs in a branch, while c_{jk} denotes the product of fixed parameter values along each branch (Hu & Batchelder, 1994).

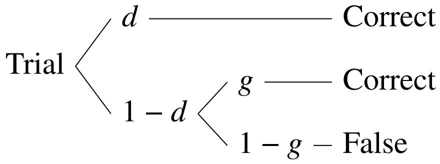

Figure 1. A simple multinomial processing tree model for a perception experiment with a two-alternative forced-choice test. d = probability to detect the stimulus; g = probability to guess correctly.

The probability of each category k as a function of the model parameters is the sum of all branch probabilities ending in category k,

$$p_k(\mathbf{\Theta}) = \sum_{j=1}^{J} p_{jk}(\mathbf{\Theta}) \, . \tag{16}$$

For the observed category frequencies $n^K = (n_1, ..., n_K)$, $\sum_{k=1}^{K} n_k = N$, the resulting likelihood function is then given by

$$\mathcal{L}(\mathbf{\Theta};\ n^K) = N! \prod_{k=1}^{K} \frac{[p_k(\mathbf{\Theta})]^{n_k}}{n_k!} \, . \tag{17}$$

For parameter interpretation, any statistical modeling requires the model fitted to the data to be identifiable. In case of an MPT model, this means that $\mathbf{\Theta} \neq \mathbf{\Theta}'$ implies that $\mathbf{p}(\mathbf{\Theta}) \neq \mathbf{p}(\mathbf{\Theta}')$, for all $\mathbf{\Theta}, \mathbf{\Theta}' \in \mathbf{\Omega}$. In other words, a model is globally identifiable if any specific set of model-consistent category probabilities corresponds to a unique set of parameter values (Bamber & van Santen, 2000).

In our example, the MPT model contains two parameters: $\mathbf{\Theta} = (d,\ g)$. In a balanced and completely randomized design, it is reasonable to assume that guessing in a 2AFC cannot be systematically "biased" towards a correct or incorrect response. Therefore, we fix the guessing parameter a priori, $g = .50$. Thus, according to (16) the probability of a correct response is given by

$$p_c(d) = d + (1 - d) \cdot .50 \, , \tag{18}$$

while the probability of an incorrect response is given by $1 - p_c(d)$, since there are only two observed response categories. The restricted model is identifiable, but since $K' = S'$, with K' denoting the number of independent categories and S' the number of free parameters, it is saturated and does not allow for tests of goodness of fit. It is still possible, however, to test hypotheses about free parameters in a saturated model.

To assess the participant's ability to detect the visual stimulus, we want to test the following hypotheses on the detection parameter d in our MPT model: $\mathcal{H}_0$: $d = 0$ versus $\mathcal{H}_1$: $d > 0$. In other words, is the response behavior based entirely on guessing or can the participant detect the stimulus at least sometimes? To control the probabilities of decision errors, we request that the test accepts $\mathcal{H}_1$ with probability $\alpha = .05$ if $d = 0$, and $\mathcal{H}_0$ with probability $\beta \leq .05$ if $d \geq .10$. To this end, we test the simple hypothesis that $d = d_0 := 0$ versus the simple alternative that $d = d_1 := .10$.

In a classical analysis, we would sample N observations from the participant and test whether our MPT model with two restricted parameters $\mathbf{\Theta}_{R_2} = (d = 0,\ g = .50)$ fits the data worse than a model where d remains unrestricted, that is, $\mathbf{\Theta}_{R_1} = (d,\ g = .50)$. As the two models are nested, the test is based on the difference of the respective fit statistics, $\Delta\text{PD}^{\lambda}$, where PD^{λ} denotes any power divergence statistic defined by λ, for example, the log-likelihood ratio statistic G^2 if $\lambda = 0$ or Pearson's χ^2 statistic if $\lambda = 1$. Under the null hypothesis defined above, $\Delta\text{PD}^{\lambda} \sim \chi^2(1)$ holds asymptotically, irrespective of the PD^{λ} statistic chosen (Read & Cressie, 1988). Thus, if $P(\chi^2(1) \geq \Delta\text{PD}^{\lambda}) < \alpha$, we decide in favor of the hypothesis $d \geq .10$.

In this example, a power analysis is straightforward. The models under $\mathcal{H}_0$ and $\mathcal{H}_1$ imply certain category probabilities. A common standardized effect size measure for the discrepancy between expected proportions under two hypotheses is Cohen's w (Cohen, 1992). In a single-tree MPT model, w is given by

$$w = \sqrt{\sum_{k=1}^{K} \frac{(p_{1k} - p_{0k})^2}{p_{0k}}}\,, \tag{19}$$

where p_{ik} denotes the probability of category k under hypothesis $i = 0, 1$. Based on (18) and (19), the expected effect size in our example with $d_0 = 0$ and $d_1 = .10$ is $w = 0.10$, denoting a small effect. Thus, a one-tailed asymptotic test of the hypothesis that $d = d_0$ versus $d = d_1$ with error probabilities $\alpha = \beta = .05$ requires approximately $N = 1{,}083$ observations (Faul et al., 2009).

Since we are dealing with simple hypotheses, the SPRT provides a most efficient alternative. Let $p_i = p_c(d_i)$, then the likelihood given hypothesis i, according to (17), is

$$\mathcal{L}(d_i;\, n_c) = \binom{N}{n_c} p_i^{n_c} (1 - p_i)^{N-n_c}\,, \tag{20}$$

where n_c denotes the observed number of correct responses. Thus, our hypotheses can be tested by means of an SPRT by continuing to sample observations from the participant as long as

$$\frac{\beta}{1-\alpha} < \frac{p_1^{n_c}(1-p_1)^{N-n_c}}{p_0^{n_c}(1-p_0)^{N-n_c}} < \frac{1-\beta}{\alpha} \tag{21}$$

and terminating as soon as one of the inequalities is violated, thus accepting either $\mathcal{H}_0$ or $\mathcal{H}_1$.

Based on formulae derived by Wald (1947), it is straightforward to approximate functions describing the test procedure's properties (see Appendix B for details). Specifically, we can analytically determine the procedure's probability to accept the alternative hypothesis (the so-called Operating Characteristic, OC) as well as the expected sample size at termination (the so-called Average Sample Number, ASN) as a function of the true value of the parameter d. The respective functions of the SPRT in this example are depicted in Figure 2. Additionally, we simulated the SPRT for the given hypotheses to demonstrate how well the procedure's properties are approximated in practice[1]. The results are denoted by the grey dots in Figure 2. Except for a slight underestimation of the ASN when the true value lies between d_0 and d_1, the analytical functions approximate the simulated estimates almost perfectly.

As the results show, the SPRT not only controls error probabilities as accurately as Neyman-Pearson tests do, it does so notably more efficiently. For any true value d, the expected sample size at termination is substantially smaller than the sample size determined by an a priori power analysis for the given hypotheses ($N = 1{,}083$). When d equals d_0 or d_1, the expected sample size of the SPRT is approximately $N = 545$, that is, almost 50% smaller. Moreover, if the true parameter value is notably larger than specified by the hypothesis, the test will require even lower sample sizes to make a decision. Classical analysis, in contrast, requires the a priori defined sample size irrespective of the true value. Thus, for the test of a simple hypothesis in a single-parameter MPT model, the SPRT is a highly efficient alternative to classical inference procedures.

3.2 Case 2: Composite Hypothesis with a Single Nuisance Parameter

In practical applications of cognitive psychometrics as well as in experimental settings, parameter tests in MPT models will rarely be on absolute parameter values as in Case 1. It is much more common to test equality or order constraints on model parameters to compare cognitive processes under different conditions or with different stimulus material. The challenge with this kind of parameter tests, however, is that they typically imply tests of composite hypotheses.

Consider the following extension of the simple psychometric experiment introduced in Case 1. Instead of the absolute perceptual ability, we

[1] R scripts for this and all following simulations as well as all simulated raw data are available from https://osf.io/98erb/.

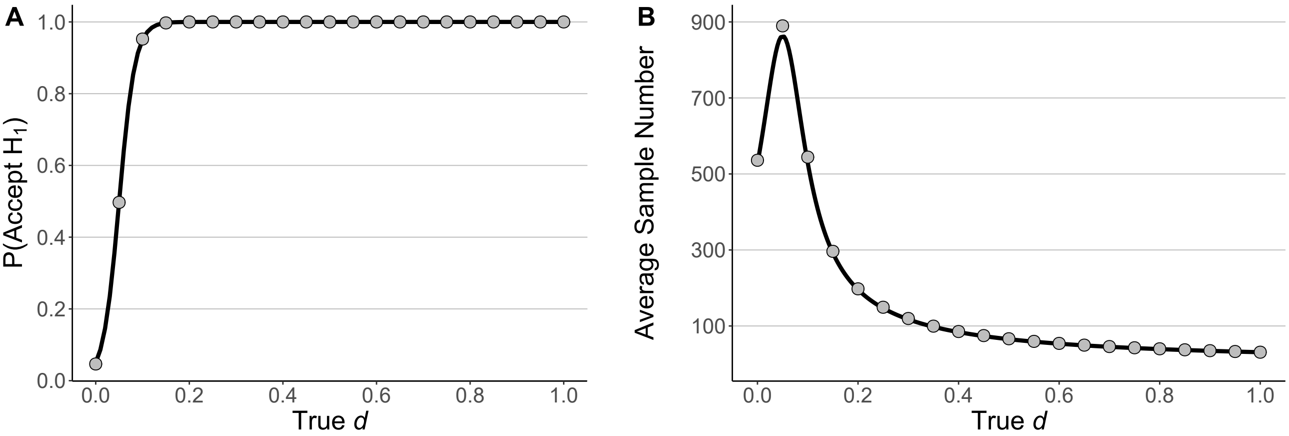

Figure 2. **A**: The black line denotes the Operating Characteristic (OC) function for a sequential probability ratio test (SPRT) on $\mathcal{H}_0$: $d = 0$ versus $\mathcal{H}_1$: $d = .10$ with $\alpha = \beta = .05$. **B**: The black line denotes the Average Sample Number (ASN) function of the respective SPRT. Grey dots denote simulated estimates of OC and ASN for the given test, based on 10,000 replications per estimate.

now want to assess the testee's perceptual sensitivity. Specifically, we manipulate physical features of the visual stimulus presented and assess whether the participant's ability to detect the stimulus differs between conditions (see Blackwell et al., 1954, for a similar experimental procedure). To this end, the stimulus is now presented in two different magnitudes (low versus high luminous intensity). As in Case 1, we want to test the detection processes directly by means of an MPT analysis of the individual's performance in the 2AFC.

Figure 3 depicts the extended MPT model for Case 2. The model now comprises two processing trees, one for each stimulus magnitude. We still assume unbiased guessing of the correct response in the 2AFC, that is, $g = .50$ for each stimulus type. However, to test whether the manipulation of stimulus magnitude affects the detection probability, the model now contains two detection parameters, d_h (high magnitude) and d_l (low magnitude). We want to test the hypotheses $\mathcal{H}_0$: $d_h = d_l$ versus $\mathcal{H}_1$: $d_h > d_l$, as the probability to detect the stimulus should be higher for high stimulus magnitude than for low magnitude.

To incorporate parametric order constraints into a binary MPT model, it is straightforward to reparameterize the model such that the new model satisfies all assumptions of binary MPT models and is statistically equivalent to the original model (Knapp & Batchelder, 2004): By restructuring the processing tree for low-intensity stimuli and introducing a new parameter ξ (Figure 4), we can express d_l in terms of d_h:

$$d_l = \xi d_h \, . \tag{22}$$

The reparameterized model, just as the original model, contains two unknown parameters, $\mathbf{\Theta} = (d_h, \ \xi)$, both of which are free to vary in the entire parameter space $\Omega = [0, \ 1]$. Our hypotheses are then reformulated in terms of ξ, that is, $\mathcal{H}_0$: $\xi = \xi_0$ $(\xi_0 = 1)$ and $\mathcal{H}_1$: $\xi = \xi_1$ $(\xi_1 < 1)$. Thus, our hypotheses are about the ratio of detection probabilities for low and high stimulus magnitude.

It is easy to see that these hypotheses are composite as the probability distribution of our data depends both on ξ, which is specified by the hypotheses, and d_h, which is unknown. This is a particular problem for a Neyman-Pearson test of the hypotheses, as the effect size and, in turn, the power of the test also depend on both parameters.

If an MPT model includes more than one tree, the model becomes a joint MPT model. For $T > 1$

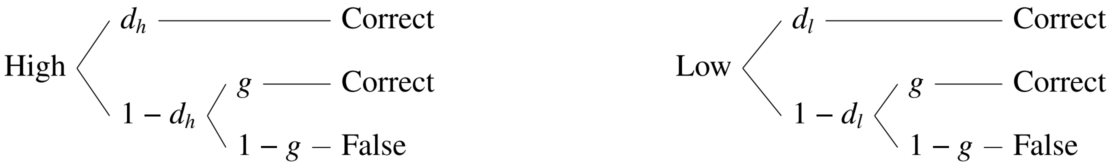

Figure 3. A multinomial processing tree model for a perception experiment with two stimulus magnitudes (high versus low luminous intensity) and a two-alternative forced-choice test. d_h = probability to detect the stimulus with high magnitude; d_l = probability to detect the stimulus with low magnitude; g = probability to guess correctly.

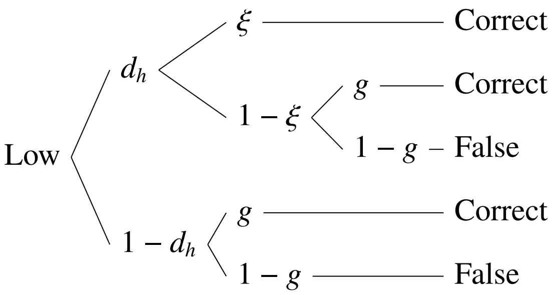

Figure 4. Reparameterization of the second processing tree depicted in Figure 3 for the order constraint $d_h > d_l$. d_h = probability to detect the stimulus with high magnitude; ξ = ratio of the probability to detect the stimulus with low magnitude to d_h; g = probability to guess correctly.

trees, Cohen's effect size measure w generalizes to

$$w = \sqrt{\sum_{t=1}^{T} \pi_t \cdot \sum_{k_t=1}^{K_t} \frac{(p_{1kt} - p_{0kt})^2}{p_{0kt}}}\,, \tag{23}$$

where K_t denotes the total number of categories in tree t ($t = 1, ..., T$) and π_t the proportion of the total sample size N assigned to tree t. Resembling Case 1, p_{1kt} denotes the predicted category probabilities for category k of tree t according to $\mathcal{H}_1$. However, since $\mathcal{H}_0$ is composite, the corresponding p_{0kt} category probabilities are now obtained by fitting the $\mathcal{H}_0$ model (with d_h free) to these $\mathcal{H}_1$ probabilities such that w becomes a minimum (Erdfelder, Faul, & Buchner, 2005). Note that (23) reduces to (19) iff $T = 1$.

Assume $\xi_0 = 1.00$ and $\xi_1 = .75$. Then the expected effect size for $d_h = .70$ in a balanced design with $\pi_{high} = \pi_{low} = .50$ is approximately $w = 0.11$ according to (23). An a priori power analysis for this effect size reveals a required sample size of $N = 892$ observations for a one-tailed asymptotic test with $\alpha = \beta = .05$. If, however, $d_h = .50$, the expected effect size is only $w = 0.07$ and the required sample size for the same test is more than twice as large, that is, $N = 2,248$.

To ensure a sufficiently powered test in the context of a composite hypothesis, a rational strategy would be to assume a conservative value of d_h such that the resulting test has power $1 - \beta \geq .95$ for any d_h in a reasonable range. However, this can be inefficient and demand very large sample sizes.

Instead, we can analyze the data sequentially by means of the SMLRT. Let $p_h = d_h \cdot (1 - d_h) \cdot .50$ denote the probability of a correct response in a trial with high stimulus magnitude under both hypotheses, and $p_{li} = \xi_i d_h + (1 - \xi_i d_h) \cdot .50$ the corresponding probability for low stimulus magnitude under hypothesis i. The likelihood function is then given by

$$\begin{aligned}&\mathcal{L}(\xi_i, d_h;\ n_1, n_2, n_3, n_4)\\ &\quad= \frac{N!}{\prod_{k=1}^{4} n_k!} p_h^{n_1} (1 - p_h)^{n_2} p_{li}^{n_3} (1 - p_{li})^{n_4}\,,\end{aligned} \tag{24}$$

where the n_k ($k = 1, 2, 3, 4$; $\sum_{k=1}^{4} n_k = N$) denote observed frequencies of correct versus false responses for high versus low stimulus magnitude, respectively.

If Equation (10) is satisfied in our case (with $\theta = \xi$ and $\phi = d_h$), that is, if $\hat{d}_h$ and $\hat{\xi}$ are asymptotically independent, the SMLRT reduces to a simple SPRT where d_h is replaced by $\hat{d}_h$ at each step. However, since

$$\frac{\partial^2 \ell(\xi, d_h;\ n_1, n_2, n_3, n_4)}{\partial\xi\partial d_h} = \frac{n_3}{(\xi d_h + 1)^2} - \frac{n_4}{(\xi d_h - 1)^2} \tag{25}$$

and considering that in a balanced design, $E(n_3) = N/2 \cdot [\xi d_h + (1 - \xi d_h) \cdot .50]$ and $E(n_4) = N/2 - E(n_3)$, we see that

$$E\left[\frac{1}{N}\frac{\partial^2 \ell(\xi, d_h;\ n_1, n_2, n_3, n_4)}{\partial\xi\partial d_h}\right] = -\frac{\xi d_h}{2(1 - \xi^2 d_h^2)}\ . \tag{26}$$

The term in (26) no longer depends on N and, thus, (10) is not satisfied if neither ξ nor d_h are equal to 0. Hence, as suggested by D. R. Cox (1963), the test should be based on (12) with stooping boundaries (13), where θ is replaced by ξ. To calculate the expected Fisher information in order to obtain $\mathcal{V}_{\xi\xi}$ at each step, observed cell frequencies in the Hessian matrix $\mathbf{H}(\xi, d_h)$ are replaced by the expected cell frequencies, as was done in (26). Additionally, when $\xi_1 < \xi_0$, the inequalities in (13) must be inverted. This will be the case for order constraints in MPT models, where the null hypothesis typically denotes $\xi_0 = 1$, such as in our example.

Unlike in the SPRT for simple hypotheses, there are no analytical formulae for the SMLRT's properties for composite hypotheses. Therefore, we simulated the SMLRT for the perception experiment in Case 2 (1) to assess whether long-run error rate control works as expected and (2) to compare the expected sample size required by the SMLRT to that of the classical Neyman-Pearson test.

The simulations were carried out in the statistical computing environment R (R Core Team, 2019). We generated participants' responses according to the model depicted in Figures 3 and 4 and analyzed them sequentially by means of the SMLRT defined by (12) and (13) with inverted boundaries. Estimates of ξ were computed with the R package MPTinR (Singmann & Kellen, 2013).

We simulated data for different true values of d_h ($d_h = .70, .50$) and ξ ($\xi = 1.00, .75, .50$). Under the null hypothesis, ξ_0 was always equal to 1, while under the alternative hypothesis, ξ_1 was equal to .75 or .50. Furthermore, we varied the initial sample size of the sequential procedure (N_{min}). As the SMLRT is based on large-sample approximations, a too small sample size might negatively affect the procedure and compromise its error rates (C. P. Cox & Roseberry, 1966). As a simple strategy to find a suitable number, the initial sample was therefore defined to be 25%, 40%, or 50% of the sample size required by a corresponding Neyman-Pearson test (N_{NP}).[2]

In each step, the sample size was increased by +2, one observation for each stimulus magnitude, until a threshold was reached. Threshold values were chosen such that $\alpha = \beta = .05$. For each parameter combination, we replicated the test procedure 1,000 times.

The results are displayed in Figure 5. It contains the empirical error rates (α' and β') and the required sample sizes as a function of d_h, the true value of ξ, that is, $\xi = \xi_0$ or $\xi = \xi_1$, and the initial sample size. Error bars for the error rates denote 95% exact confidence intervals (Clopper & Pearson, 1934). The sample size distributions are displayed as boxplots. Black dots denote outliers (data points further than 1.5 times the inter-quartile range below or above the first or the third quartile, respectively), grey dots represent the means of the distributions, that is, the ASN. Dashed lines denote the nominal error rates and the sample sizes required by a corresponding Neyman-Pearson test.

The left part of Figure 5 shows the results for $\xi_1 = .50$, the right part displays results for $\xi_1 = .75$.

[2] To increase computational efficiency, each simulated trajectory started with $N_{min} = .25 \cdot N_{NP}$ and was then reanalyzed with $N_{min} = .40 \cdot N_{NP}$ and $N_{min} = .50 \cdot N_{NP}$.

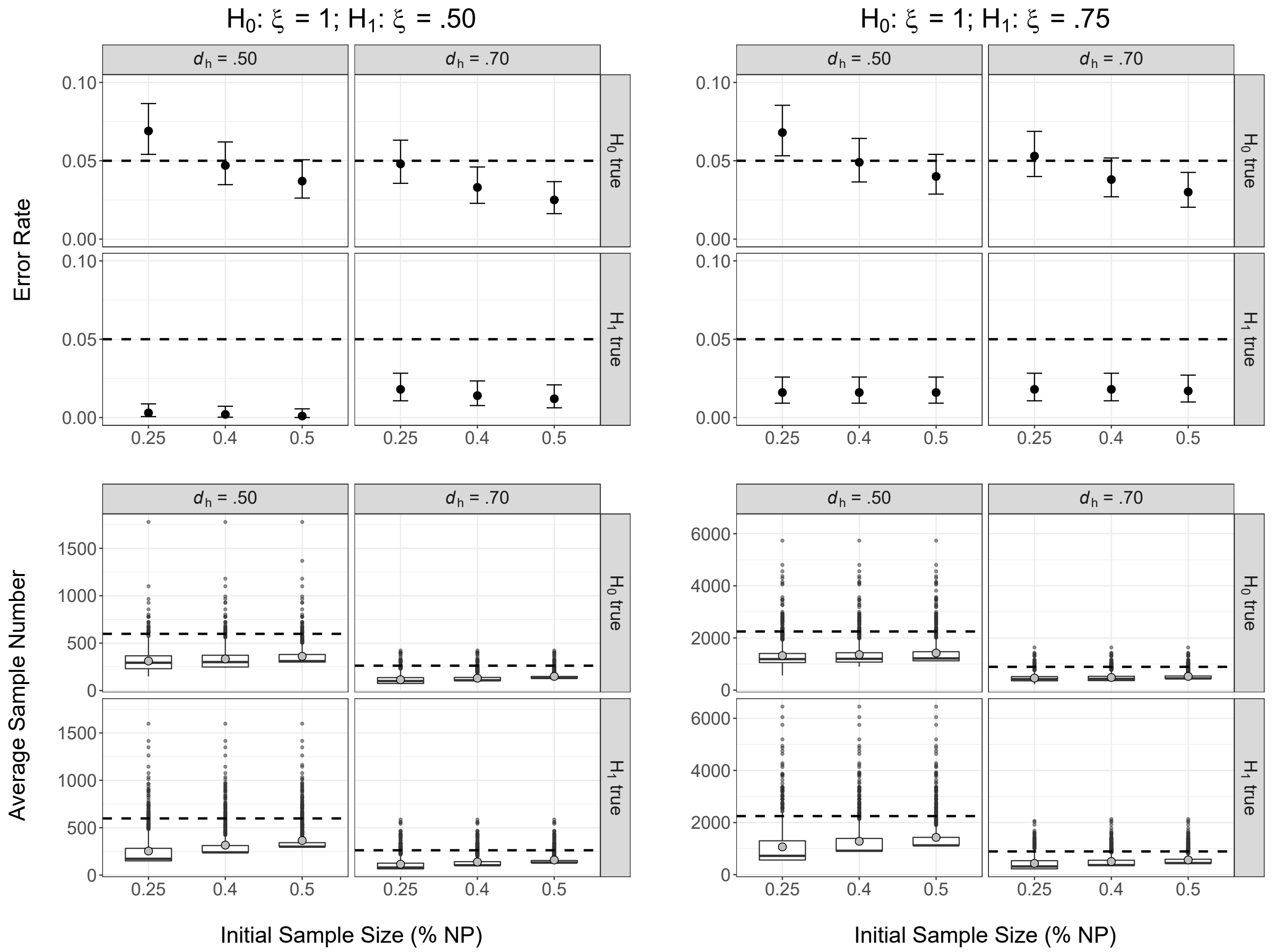

Figure 5. Empirical error rates and sample size distributions of the sequential maximum likelihood ratio test (SMLRT) as a function of the hypothesis tested, the true detection parameter d_h, and the data-generating scenario. Error bars denote 95% Clopper-Pearson exact confidence intervals (Clopper & Pearson, 1934). Black dots in the boxplots denote outliers (data points more than 1.5 times the inter-quartile range below or above 1st or 3rd quantile). Grey dots denote mean sample sizes. Dashed lines represent nominal error rates and sample sizes required by a corresponding Neyman-Pearson (NP) test.

For all parameter combinations, β' substantially undercuts the nominal level. At the same time, except for a slight upward deviation when $d_h = .50$ and the initial sample size is small, α' adheres to the nominal level. Moreover, the SMLRT controls error probabilities notably more efficiently than a corresponding Neyman-Pearson test: The ASN is on average 45% smaller. Across all parameter combinations, the test terminates with a sample size smaller than N_{NP} in 94% of the cases.

In almost all of the simulated scenarios, the SMLRT shows satisfying results for an initial sample size of $N_{min} = .25 \cdot N_{NP}$. With increasing N_{min}, the test procedure becomes more conservative and less efficient. However, the increase in ASN is only slight and still below the sample size required by the Neyman-Person test. Concluding from our results, an initial sample size of 25% of a corresponding Neyman-Pearson test is a reasonable starting point to efficiently control long-run rates of statistical decision errors for parameter tests in MPT models with a single unknown nuisance parameter.

3.3 Case 3: Composite Hypothesis with Several Nuisance Parameters

Commonly, MPT models contain several unknown parameters. Thus, hypotheses about single parameters typically involve more than one nuisance parameter. To illustrate that the SMLRT naturally extends to this case, consider the following variation of our psychometric experiment.

To assess potential biases involved in the decision process as well as perceptual processes, the experiment is now based on a Yes/No test. That is, in each trial either a stimulus (target) or no stimulus (lure) is presented. For each trial, the participant has to indicate whether they detected a stimulus ("Yes") or not ("No"). As in Case 2, stimuli are light flashes presented in two different magnitudes (high versus low luminous intensity).

Figure 6 displays the MPT model for Case 3. It contains three detection parameters denoting the probability to detect a stimulus with high magnitude (d_h), a stimulus with low magnitude (d_l), or the absence of a stimulus (d_n). Additionally, it contains the parameter g, which represents the conditional probability to guess "Yes" in a state of uncertainty.

With $K^{'} < S^{'}$, the model is not identifiable. Thus, we need to restrict at least one of the parameters. As g no longer refers to guessing correctly but rather guessing that a stimulus was presented, it seems reasonable not to restrict it a priori. For the given experiment, we rather assume that the absence of a stimulus should be equally salient and detectable as the presence of a high-magnitude stimulus. Thus, we will assume that $d_n = d_h$. The restricted model is identifiable and saturated.

To test whether the participant is sensitive to the manipulation of stimulus magnitude in the new paradigm, we will reparameterize the model as we did in Case 2 (see Figure 4), such that $d_l = \xi d_h$. Again, we test the hypotheses $\mathcal{H}_0$: $\xi = \xi_0$ ($\xi_0 = 1$) versus $\mathcal{H}_1$: $\xi = \xi_1$ ($\xi_1 < 1$). This time, the power of a hypothesis test on ξ not only depends on d_h but also on the bias to respond "Yes", g.

Similar to Case 2, the effect size for this case can be calculated based on (23), this time with $T = 3$ and $\pi_{high} = \pi_{low} = \pi_{lure} = .33$. When testing $\xi_0 = 1.00$ versus $\xi_1 = .75$ while assuming $d_h = .70$ and $g = .50$ for the nuisance parameters under $\mathcal{H}_1$ (while treating d_h and g as free parameters under $\mathcal{H}_0$), the effect size is $w = 0.09$. A classical one-tailed asymptotic test with $\alpha = \beta = .05$ would thus require $N = 1,335$ observations. However, if the participant has a slight bias to respond with "Yes" under uncertainty, $g = .60$, the effect size is reduced to $w = 0.08$ and the same test would require about $N = 1,752$ observations. For $g = .40$, in contrast, the required sample size reduces to about $N = 1,059$. Also taking into account different possible values of d_h would further increase the number of possible power analyses, thus illustrating the difficulty of determining a reasonable sample size for classical hypothesis tests in MPT models with more than one unknown nuisance parameter.

In the SMLRT, in contrast, we only need $\hat{\xi}$ and $\mathcal{V}_{\xi\xi}$ (or $SE_{\hat{\xi}}$), as the test is based on (12) and (13). The uncertainty with respect to the actual values of the nuisance parameters is taken into account implicitly, since $\mathcal{V}_{\xi\xi}$ and, correspondingly, the standard error of $\hat{\xi}$ depend in general on the precision and values of all parameter estimates. In the same vein, further increasing the complexity of the model in terms of the number of nuisance parameters or experimental conditions would not alter the general procedure for testing hypotheses on ξ.

As shown in the previous simulation, however, the SMLRT requires a sufficiently large initial sample size. If the sample size is too low, error rates may be inflated. If it is too large, the test may be less efficient. To illustrate that the required initial sample size may depend on the values of the nuisance parameters, we simulated the SMLRT for Case 3. The settings of the simulation were essentially identical to those in the previous simulation. Additionally, we varied the guessing parameter $g = .40$ versus $g = .60$. As the experiment in Case 3 comprises three stimulus categories (high

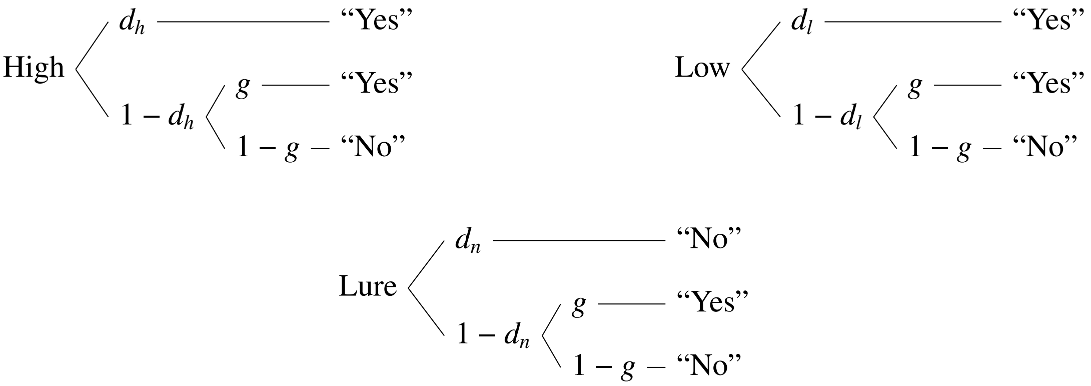

Figure 6. A multinomial processing tree (MPT) model for a perception experiment with two stimulus magnitudes (high versus low) and a Yes/No test. d_h = probability to detect the stimulus with high magnitude; d_l = probability to detect the stimulus with low magnitude; d_n = probability to detect a lure trial in which no stimulus was presented; g = probability to guess "Yes".

versus low magnitude targets and lures), the sample size was increased by +3 in each step, one observation per stimulus category. For each parameter combination, $1,000$ replications were simulated.

The results are displayed in Figure 7. For all simulated parameter combinations, the test shows very low rates of Type 1 errors. At the same time, however, the ASN in this case is still on average 38% smaller than the Neyman-Pearson sample size. The empirical β' closely approximate the nominal error rate for almost all parameter combinations. Only when $d_h = .70$ and $g = .60$, the test of $\xi_1 = .50$ yields too large β' when the initial sample size is smaller than $.50 \cdot N_{NP}$. Across all parameter combinations, the SMLRT is on average 34% more efficient than a Neyman-Pearson test and terminates with a smaller sample in 88% of the cases.

As our simulations show, the general procedure of the SMLRT extends to models with more than one unknown nuisance parameter. However, we also see the importance of a sufficiently large initial sample size in this case. When both d_h and g are large, the model predicts very low probabilities of "No" responses. In case of a large expected effect such as $\xi_1 = .50$, the classical Neyman-Pearson test is already quite efficient. Consequently, an initial sample based on 25% or 40% of N_{NP} is so small that the risk of extremely small cell frequencies is high. In such a case, the asymptotic approximations upon which the SMLRT is based cannot be expected to hold (cf. C. P. Cox & Roseberry, 1966).

For example, if $d_h = .70$ and $g = .60$, a classical one-tailed test requires $N = 174$ observations per tree to test $\xi_0 = 1$ versus $\xi_1 = .50$ with $\alpha = \beta = .05$. Thus, an initial sample size of 25% of the Neyman-Pearson sample size would comprise $N = 44$ observations per tree only. Conditional on the assumed values of d_h and g, the expected number of incorrect responses for high-magnitude targets in this case is only $44 \cdot (1 - .70) \cdot (1 - .60) = 5.28$. Not surprisingly, the large-sample approximations on which the SMLRT is based do not hold in such a situation. This could be remedied by further increasing the initial sample size. In that case, however, the test would no longer be more efficient than a classical test procedure.

It is important to note, however, that a case in which the classical test is already so efficient that the SMLRT cannot satisfy the nominal error rates with smaller samples is of practical relevance only if we can place high confidence in the parameter assumptions we make. Under uncertainty, we would rather rely on conservative assumptions to

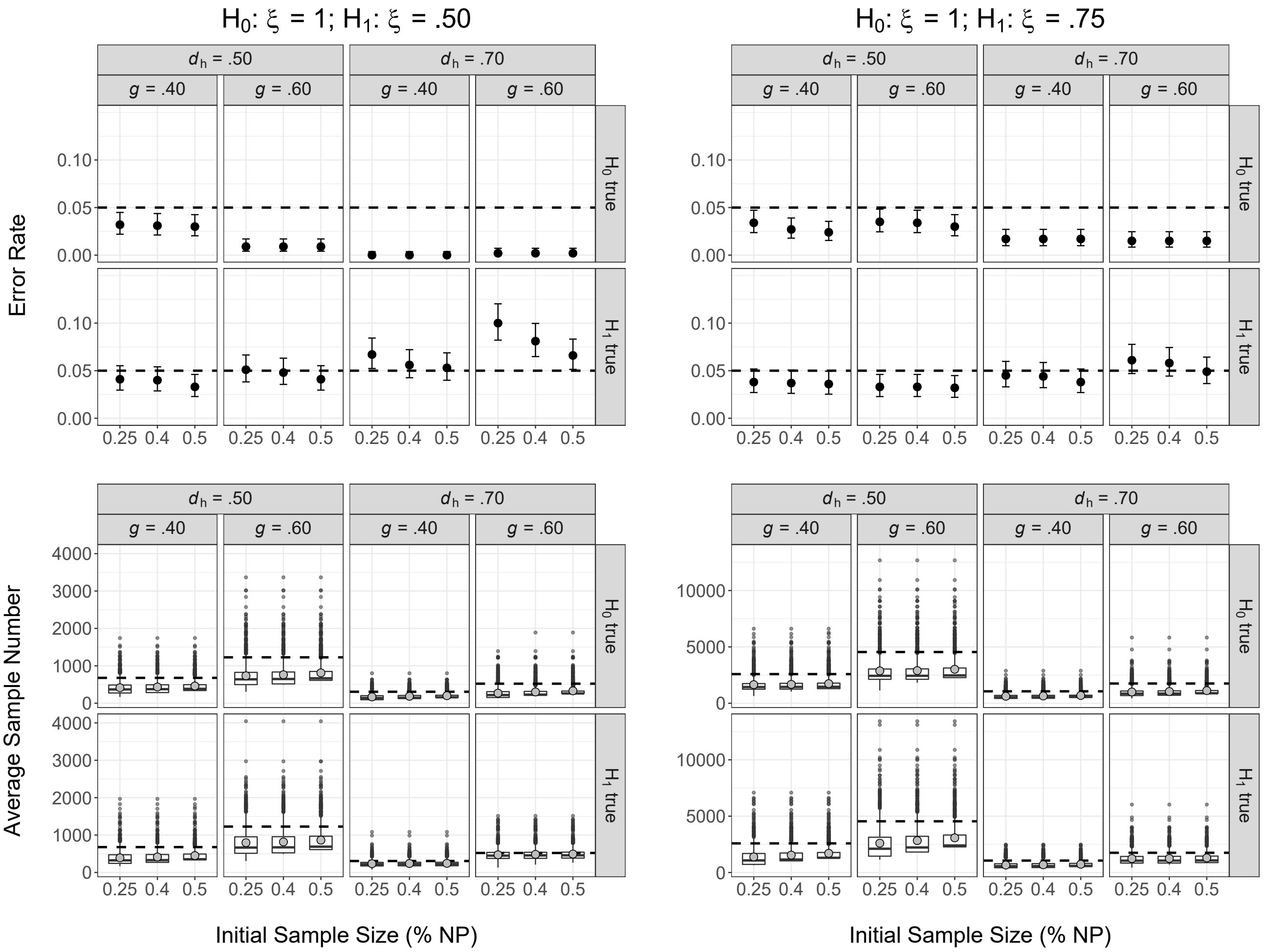

Figure 7. Empirical error rates and sample size distributions of the sequential maximum likelihood ratio test as a function of the hypothesis tested, the true detection parameter d_h, the true guessing parameter g, and the data-generating scenario. Error bars denote 95% Clopper-Pearson exact confidence intervals (Clopper & Pearson, 1934). Black dots in the boxplots denote outliers (data points more than 1.5 times the inter-quartile range below or above 1st or 3rd quantile). Grey dots denote mean sample sizes. Dashed lines represent nominal error rates and sample sizes required by a corresponding Neyman-Pearson (NP) test.

ensure sufficient power. If we follow this advice, the SMLRT will in general be more efficient.

4 Discussion

Hypothesis tests on parameter constraints in MPT models often rely on NHST, thus ignoring statistical power. Although power analyses have been worked out for categorical data (Erdfelder et al., 2005) and are readily available in existing software (e.g., Faul et al., 2009; Moshagen, 2010), practitioners typically face two challenges. First, to determine the effect size for a hypothesis test on a single parameter in a multi-parameter MPT (or other) model, the population values of all other parameters must be known or specified a priori based on theoretical considerations that may or may not hold. As a remedy, one can perform multiple power analyses for a range of reasonable parameter values and then choose the most conservative one. However, this strategy often fosters a second challenge for practitioners, namely, that the required sample sizes may become very large and practically infeasible.

As a remedy, in the present article we suggest to rely on sequential tests, an efficient alternative to classical statistical methods for hypothesis tests in MPT models. Sequential hypothesis tests control error probabilities of statistical decisions just as classical Neyman-Pearson tests do. Yet, at the same time, they are based on continuous monitoring of the data as they are sampled and terminate as soon as the data contain sufficient evidence for one hypothesis vis-à-vis the other. Thus, on average, sequential tests require notably smaller samples than classical methods that are based on a priori defined sample sizes (Schnuerch & Erdfelder, 2019).

We introduced the SPRT (Wald, 1947) and demonstrated how it is easily applied to analysis of MPT models with a single free parameter. We showed that it is substantially more efficient than classical Neyman-Pearson tests, requiring about 50% smaller samples on average. However, although there are applications of single-parameter MPT models in the literature (e.g., models for the randomized response technique; see Ulrich, Schröter, Striegel, & Simon, 2012), MPT models that are commonly used in cognitive psychology typically contain more than one parameter, many of which are nuisance parameters that need to be estimated.

Therefore, we introduced an extension of the SPRT suggested by D. R. Cox (1963) for sequential tests of composite hypotheses. In the SMLRT, the likelihood ratio is constructed based on ML estimates of both the test-relevant and the nuisance parameters. The sequential procedure is then corrected for the additional estimation uncertainty, such that the resulting test does not exceed long-run error rates α and β. Hence, the test procedure controls error probabilities without requiring knowledge or a specification of the exact values for the unknown nuisance parameters in the statistical model.

We illustrated how the SMLRT can be used to test hypotheses on MPT model parameters with existing MPT software. Essentially, the procedure merely requires the ML estimate $\hat{\theta}$ of the test-relevant parameter and the expected Fisher information (or the standard error of the estimate). Moreover, the SMLRT does not only remedy the problem of unknown nuisance parameters, it also increases efficiency of hypothesis testing. We demonstrated by means of simulations that the SMLRT requires on average 34% (Case 3) to 45% (Case 2) smaller samples to satisfy the same or even lower error rates compared to classical Neyman-Pearson tests even when these are based on the true, data-generating values of the nuisance parameter (which is an unlikely assumption in practice).

The sequential approach can be particularly useful in individual assessments (e.g., clinical diagnosis). As part of Bill Batchelder's proposal of cognitive psychometrics—that is, building a bridge between the fields of mathematical psychology and psychometrics—he strongly promoted the use of MPT models in the context of psychological assessment (Batchelder, 1998). For instance, he identified the great potential for substantive MPT model applications as diagnostic tools in clinical settings (Batchelder & Riefer, 1999; Riefer, Knapp, Batchelder, Bamber, & Manifold, 2002). However, Batchelder also acknowledged the obvious drawback of reduced estimation precision and low statistical power as a consequence of the small number of data points on the individual level. The sequential approach we promote in this article may facilitate the application of MPT models in individual assessments whenever it is necessary to make decisions about the presence or absence of specific cognitive symptoms while controlling error probabilities. More generally, we hope that the SMLRT for MPT models will further contribute to the increasing number of substantive applications in cognitive psychometrics.

Apart from individual assessment, sequential analysis is also particularly useful for efficient MPT modeling of data on the group level when each participant provides only a single data point (e.g., Heck, Thielmann, et al., 2018; Klauer et al., 2007; Moshagen et al., 2014; Moshagen et al., 2012; Schild et al., 2019). This setting has the advantage that the MPT analysis need not be built into the experimental procedure or software because the data can be analyzed after data collection for each participant. It is thus easily implemented in practice, thereby providing an attractive alternative to classical methods in terms of a more efficient and less costly control of error probabilities.

4.1 Limitations

The approaches presented in this article are so-called unrestricted sequential procedures that do not have a definite upper bound of sample size. Hence, although the test is on average more efficient than classical procedures, there is a potential risk that the data provide inconclusive evidence in single cases, meaning that the test will continue for a long time without reaching one of the two boundaries. Concluding from our simulation results, this risk is small (approximately 6% in Case 2, 12% in Case 3). Nevertheless, this risk potentially limits its applicability in individual analysis to situations in which the number of data points is not restricted a priori. Think, for example, of an experimental paradigm assessing long-term episodic memory processes (e.g., Batchelder & Riefer, 1986). Such a paradigm typically includes a learning phase and a test phase. The number of possible data points in the test phase is limited by the number of items learned during the first phase. Thus, sequential analysis during the test phase will make sense only if it requires no more than the number of learned items. As this obviously cannot be guaranteed, the unrestricted sequential approach is not appropriate for such applications.

Second, the SMLRT for composite hypotheses is based on large-sample approximations (D. R. Cox, 1963). Therefore, as the simulation results in Case 3 showed, the method may fail when initial sample sizes are too small (see also C. P. Cox & Roseberry, 1966; Wetherill, 1975). The relevance of a sufficiently large initial sample size increases with model complexity, as does the required sample size at termination. The practical challenge is of course to determine a suitable initial sample size for the sequential procedure on a priori grounds. If the sample size is too small, error rates might be seriously inflated. If it is too large, on the other hand, the test's efficiency is reduced (although our simulations demonstrated that the increase in ASN due to larger initial sample sizes is only slight).

As a remedy, we suggest to search for a model-specific minimum sample size by means of an a priori power analysis and Monte Carlo simulations. Of course, this will again entail assumptions about reasonable true parameter values of the nuisance parameters. However, the consequences of an overly conservative assumption in the context of a sequential test are much less severe than for a standard test procedure. If the initial sample size is chosen too large, the evidence provided by the data may already be compelling very early during data collection, meaning that the test procedure will stop immediately. Thus, the SMLRT will be more efficient than a correspondingly conservative classical test, in which one cannot use optional stopping even if the data clearly speak in favor of one of the hypotheses.

Third, sequential approaches assume that observations are independent and identically distributed (i.i.d.). This assumption is reasonable for sequential analyses of data generated by an individual provided that the experimental design prevents contaminations of the data by exercise effects, fatigue effects, or order effects. The i.i.d. assumption is also plausible in model applications where each participant provides a single data point only. If, however, MPT models are applied to aggregate data of repeated observations of multiple individuals, the i.i.d. assumption may be questioned and is often implausible (Smith & Batchelder, 2008). If there is heterogeneity in items or participants, ignoring the hierarchical structure might bias parameter estimates and statistical tests (Heck, Arnold, & Arnold, 2018). Thus, if parameter tests are performed at the group level based on data aggregated across items and participants, the sequential approaches promoted herein may not be suitable. This issue is especially critical for sequential tests if the data are collected in a batch-wise fashion. For instance, if one first collects 100 observations from a person that does perform extraordinarily well, the sequential test may already indicate a decision, thereby ignoring data from other participants that perform worse.

Finally, it is important to keep in mind that both SPRT and SMLRT address problems of hypothesis testing, not estimation. In this article, we focused on efficiency in statistical decision making exclusively. If this is the primary concern, SPRT and SMLRT are appropriate alternatives to classical methods. However, if the aim is to estimate a parameter as precisely as possible, these sequential procedures are not suitable. Whereas efficiency requires to make a decision with as few observations as possible, high precision of parameter estimates is achieved with as many observations as possible (without optional stopping depending on the current value of the estimates). In fact, parameter estimates following a sequential hypothesis test may be biased (Whitehead, 1986). Thus, the sequential approach promoted herein should only be used if the aim is in fact to make an efficient statistical decision, for example, in psychological assessments.

4.2 Conclusion

Multinomial processing tree models have proven useful in many areas of cognitive and social psychology as tools to measure and disentangle latent cognitive processes. As repeatedly argued and demonstrated by Bill Batchelder, they have great potential especially for psychometric purposes, for example, in the context of individual diagnostics in clinical settings (e.g., Batchelder, 1998; Batchelder & Riefer, 1999; Riefer et al., 2002). We introduced sequential test procedures proposed by Wald (1947) and D. R. Cox (1963) and illustrated how they can be adapted to MPT model analysis. By means of simulations, we demonstrated the excellent properties of the sequential approach for testing hypotheses on MPT model parameters both in the absence and presence of nuisance parameters. Thereby, we hope to improve efficiency of statistical inference in MPT modeling, particularly in the context of individual assessments (i.e., cognitive psychometrics) and other settings with scarce resources.

References

Bamber, D., & van Santen, J. P. H. (2000). How to assess a model's testability and identifiability. *Journal of Mathematical Psychology*, *44*, 20–40. doi:10.1006jmps.1999.1275

Barnard, G. A. (1952). The frequency justification of certain sequential tests. *Biometrika*, *39*, 144–150. doi:10.2307/2332473

Bartlett, M. S. (1946). The large-sample theory of sequential tests. *Mathematical Proceedings of the Cambridge Philosophical Society*, *42*, 239–244. doi:10.1017/S0305004100022994

Batchelder, W. H. (1998). Multinomial processing tree models and psychological assessment. *Psychological Assessment*, *10*, 331–344. doi:10.1037/1040-3590.10.4.331

Batchelder, W. H., & Riefer, D. M. (1986). The statistical analysis of a model for storage and retrieval processes in human memory. *British Journal of Mathematical and Statistical Psychology*, *39*, 129–149. doi:10.1111/j.2044-8317.1986.tb00852.x

Batchelder, W. H., & Riefer, D. M. (1990). Multinomial processing models of source monitoring. *Psychological Review*, *97*, 548–564. doi:10.1037/0033-295X.97.4.548

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review*, *6*, 57–86. doi:10.3758/BF03210812

Blackwell, H. R., Pritchard, B. S., & Ohmart, J. G. (1954). Automatic apparatus for stimulus presentation and recording in visual threshold experiments. *Journal of the Optical So-*

ciety of America, *44*, 322–326. doi:10.1364/JOSA.44.000322

Breslow, N. (1969). On large sample sequential analysis with applications to survivorship data. *Journal of Applied Probability*, *6*, 261–274. doi:10.2307/3211997

Chechile, R. A. (2009). Pooling data versus averaging model fits for some prototypical multinomial processing tree models. *Journal of Mathematical Psychology*, *53*, 562–576. doi:10.1016/j.jmp.2009.06.005

Clopper, C. J., & Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, *26*, 404–413. doi:10.2307/2331986

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159. doi:10.1037/0033-2909.112.1.155

Cox, C. P., & Roseberry, T. D. (1966). A large sample sequential test, using concomitant information, for discrimination between two composite hypotheses. *Journal of the American Statistical Association*, *61*, 357–367. doi:10.2307/2282824

Cox, D. R. (1952). Sequential tests for composite hypotheses. *Mathematical Proceedings of the Cambridge Philosophical Society*, *48*, 290–299. doi:10.1017/S030500410002764X

Cox, D. R. (1963). Large sample sequential tests for composite hypotheses. *Sankhyā: The Indian Journal of Statistics, Series A (1961-2002)*, *25*, 5–12.

Erdfelder, E., Auer, T.-S., Hilbig, B. E., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial processing tree models: A review of the literature. *Zeitschrift für Psychologie / Journal of Psychology*, *217*, 108–124. doi:10.1027/0044-3409.217.3.108

Erdfelder, E., Faul, F., & Buchner, A. (2005). Power analysis for categorical methods. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (pp. 1565–1570). Chichester, UK: Wiley.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149–1160. doi:10.3758/BRM.41.4.1149

Heck, D. W., Arnold, N. R., & Arnold, D. (2018). TreeBUGS: An R package for hierarchical multinomial-processing-tree modeling. *Behavior Research Methods*, *50*, 264–284. doi:10.3758/s13428-017-0869-7

Heck, D. W., Erdfelder, E., & Kieslich, P. J. (2018). Generalized processing tree models: Jointly modeling discrete and continuous variables. *Psychometrika*, *83*, 893–918. doi:10.1007/s11336-018-9622-0

Heck, D. W., Thielmann, I., Moshagen, M., & Hilbig, B. E. (2018). Who lies? A large-scale reanalysis linking basic personality traits to unethical decision making. *Judgment and Decision Making*, *13*, 356–371.

Hu, X. (1999). Multinomial processing tree models: An implementation. *Behavior Research Methods, Instruments, & Computers*, *31*, 689–695. doi:10.3758/BF03200747

Hu, X., & Batchelder, W. H. (1994). The statistical analysis of general processing tree models with the EM algorithm. *Psychometrika*, *59*, 21–47. doi:10.1007/BF02294263

Hu, X., & Phillips, G. A. (1999). GPT.EXE: A powerful tool for the visualization and analysis of general processing tree models. *Behavior Research Methods, Instruments, & Computers*, *31*, 220–234. doi:10.3758/BF03207714

Joanes, D. N. (1972). Sequential tests of composite hypotheses. *Biometrika*, *59*, 633–637. doi:10.1093/biomet/59.3.633

Klauer, K. C. (2006). Hierarchical multinomial processing tree models: A latent-class approach. *Psychometrika*, *71*, 7–31. doi:10.1007/s11336-004-1188-3

Klauer, K. C. (2010). Hierarchical multinomial processing tree models: A latent-trait approach. *Psychometrika*, *75*, 70–98. doi:10.1007/s11336-009-9141-0

Klauer, K. C., Stahl, C., & Erdfelder, E. (2007). The abstract selection task: New data and an almost comprehensive model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 680–703. doi:10.1037/0278-7393.33.4.680

Knapp, B. R., & Batchelder, W. H. (2004). Representing parametric order constraints in multi-trial applications of multinomial processing tree models. *Journal of Mathematical Psychology*, *48*, 215–229. doi:10.1016/j.jmp.2004.03.002

Ly, A., Marsman, M., Verhagen, J., Grasman, R. P., & Wagenmakers, E.-J. (2017). A tutorial on Fisher information. *Journal of Mathematical Psychology*, *80*, 40–55. doi:10.1016/j.jmp.2017.05.006

Matthes, T. K. (1963). On the optimality of sequential probability ratio tests. *The Annals of Mathematical Statistics*, *34*, 18–21. doi:10.1214/aoms/1177704239

Moshagen, M. (2010). multiTree: A computer program for the analysis of multinomial processing tree models. *Behavior Research Methods*, *42*, 42–54. doi:10.3758/BRM.42.1.42

Moshagen, M., Hilbig, B. E., Erdfelder, E., & Moritz, A. (2014). An experimental validation method for questioning techniques that assess sensitive issues. *Experimental Psychology*, *61*, 48–54. doi:10.1027/1618-3169/a000226

Moshagen, M., Musch, J., & Erdfelder, E. (2012). A stochastic lie detector. *Behavior Research Methods*, *44*, 222–231. doi:10.3758/s13428-011-0144-2

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*, 289–337. doi:10.1098/rsta.1933.0009

R Core Team. (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from http://www.R-project.org/

Read, T. R. C., & Cressie, N. A. C. (1988). *Goodness-of-fit statistics for discrete multivariate data*. New York, NY: Springer.

Riefer, D. M., & Batchelder, W. H. (1988). Multinomial modeling and the measurement of cognitive processes. *Psychological Review*, *95*, 318–339. doi:10.1037/0033-295X.95.3.318

Riefer, D. M., Knapp, B. R., Batchelder, W. H., Bamber, D., & Manifold, V. (2002). Cognitive psychometrics: Assessing storage and retrieval deficits in special populations with multinomial processing tree models. *Psychological Assessment*, *14*, 184–201. doi:10.1037/1040-3590.14.2.184

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

Schild, C., Heck, D. W., Ścigała, K. A., & Zettler, I. (2019). Revisiting REVISE: (Re)Testing unique and combined effects of REminding, VIsibility, and SElf-engagement manipulations on cheating behavior. *Journal of Economic Psychology*, *75*, 102161. doi:10.1016/j.joep.2019.04.001

Schnuerch, M., & Erdfelder, E. (2019). Controlling decision errors with minimal costs: The sequential probability ratio t test. *Psychological Methods.* Advance online publication. doi:10.1037/met0000234

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

Singmann, H., & Kellen, D. (2013). MPTinR: Analysis of multinomial processing tree models in R. *Behavior Research Methods*, *45*, 560–575. doi:10 . 3758 / s13428 - 012 - 0259-0

Smith, J. B., & Batchelder, W. H. (2008). Assessing individual differences in categorical data. *Psychonomic Bulletin & Review*, *15*, 713–731. doi:10.3758/PBR.15.4.713

Smith, J. B., & Batchelder, W. H. (2010). Beta-MPT: Multinomial processing tree models for addressing individual differences. *Journal of Mathematical Psychology*, *54*, 167–183. doi:10.1016/j.jmp.2009.06.007

Swets, J. A., Tanner, W. P., & Birdsall, T. G. (1961). Decision processes in perception. *Psychological Review*, *68*, 301–340. doi:10.1037/h0040547

Ulrich, R., Schröter, H., Striegel, H., & Simon, P. (2012). Asking sensitive questions: A statistical power analysis of randomized response models. *Psychological Methods*, *17*, 623–641. doi:10.1037/a0029314

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

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

Wetherill, G. B. (1975). *Sequential methods in statistics* (2. ed.). London: Chapman and Hall.

Whitehead, J. (1986). On the bias of maximum likelihood estimation following a sequential test. *Biometrika*, *73*, 573–581. doi:10.1093/biomet/73.3.573

Appendix A

Sequential Maximum Likelihood Ratio Tests

To show that (11) is asymptotically equivalent to (9), consider that according to ML theory, the expected Fisher information matrix for a sample of size n is given by

$$\begin{aligned} n\mathcal{I}(\theta,\phi) &= E\left[-\mathbf{H}(\theta,\phi)\right] \\ &= E\left[-\begin{pmatrix} \dfrac{\partial^2\ell(\theta,\phi;\ x^n)}{\partial\theta^2} & \dfrac{\partial^2\ell(\theta,\phi;\ x^n)}{\partial\theta\partial\phi} \\ \dfrac{\partial^2\ell(\theta,\phi;\ x^n)}{\partial\phi\partial\theta} & \dfrac{\partial^2\ell(\theta,\phi;\ x^n)}{\partial\phi^2} \end{pmatrix}\right] \end{aligned} \tag{27}$$

where $\mathbf{H}(\theta,\phi)$ denotes the Hessian matrix of second-order partial derivatives. Accordingly, $n\mathcal{I}_{\theta\theta}$ and $n\mathcal{I}_{\theta\phi}$ denote the (θ, θ) and (θ, ϕ) element (or submatrix) of this matrix. Moreover, $\hat{\theta}$, $\hat{\phi}$ asymptotically satisfy the following equation (D. R. Cox, 1963):

$$n\left[\mathcal{I}_{\theta\theta}(\hat{\theta}-\theta)+\mathcal{I}_{\theta\phi}(\hat{\phi}-\phi)\right] = \frac{\partial\ell(\theta,\phi;\ x^n)}{\partial\theta}\ . \tag{28}$$

Thus, writing (9) in terms of (27) and (28) gives

$$\begin{aligned} &n(\theta_1-\theta_0)\mathcal{I}_{\theta\theta}(\hat{\theta}-\theta)\ +\ n(\theta_1-\theta_0)\mathcal{I}_{\theta\phi}(\hat{\phi}-\phi) \\ &-\tfrac{1}{2}(\theta_1-\theta_0)(\theta_1+\theta_0-2\theta)n\mathcal{I}_{\theta\theta} \\ &-(\theta_1-\theta_0)(\hat{\phi}-\phi)n\mathcal{I}_{\theta\phi} \end{aligned} \tag{29}$$

which by application of simple calculus yields

$$\begin{aligned} &n(\theta_1-\theta_0)\Big[\mathcal{I}_{\theta\theta}(\hat{\theta}-\theta)+\mathcal{I}_{\theta\phi}(\hat{\phi}-\phi) \\ &\qquad -\tfrac{1}{2}\mathcal{I}_{\theta\theta}(\theta_1+\theta_0-2\theta)-\mathcal{I}_{\theta\phi}(\hat{\phi}-\phi)\Big] \\ &= n(\theta_1-\theta_0)\mathcal{I}_{\theta\theta}(\hat{\theta}-\theta\ -\tfrac{1}{2}\theta_1-\tfrac{1}{2}\theta_0+\theta) \\ &= n\mathcal{I}_{\theta\theta}(\theta_1-\theta_0)\left[\hat{\theta}-\tfrac{1}{2}(\theta_1+\theta_0)\right]. \end{aligned} \tag{30}$$

Appendix B

Properties of the Sequential Probability Ratio Test

To approximate the functions describing power and expected sample size of the sequential probability ratio test (SPRT) for a test of hypotheses about d in the MPT model displayed in Figure 1 (with g = .50), we can use formulae derived by Wald (1947). For any given d_0, d_1, α, and β, the power of the SPRT is a function of the true value

d. Let Ψ_d denote the probability to accept $\mathcal{H}_1$ given a certain true value d, then

$$\Psi_d \approx \frac{1-\left(\frac{\beta}{1-\alpha}\right)^h}{\left(\frac{1-\beta}{\alpha}\right)^h-\left(\frac{\beta}{1-\alpha}\right)^h}\ , \tag{31}$$

where h is the non-zero root of the equation

$$p\left(\frac{p_1}{p_0}\right)^h+(1-p)\left(\frac{1-p_1}{1-p_0}\right)^h=1 \tag{32}$$

with p and p_i denoting the true and predicted probability of a correct response under hypothesis i, respectively, $p_i = d_i + (1 - d_i)\cdot .50$.

It is easy to see that if $d = d_1$, which means that $p = p_1$, the non-zero root of (32) is $h = -1$,

$$\begin{aligned} &p_1\frac{p_0}{p_1}+(1-p_1)\frac{(1-p_0)}{(1-p_1)}-1\\ &=p_0+(1-p_0)-1\\ &=0\ , \end{aligned} \tag{33}$$

which, as expected, yields

$$\begin{aligned} \Psi_{d=d_1} &= \frac{1-\left(\frac{1-\alpha}{\beta}\right)}{\left(\frac{\alpha}{1-\beta}\right)-\left(\frac{1-\alpha}{\beta}\right)}\\ &= \frac{\alpha+\beta-1}{\beta}\cdot\frac{\beta(1-\beta)}{\beta\alpha-(1-\beta)(1-\alpha)}\\ &= 1-\beta. \end{aligned} \tag{34}$$

In the same vein, if $d = d_0$ the non-zero root of (32) is $h = 1$,

$$\begin{aligned} &p_0\frac{p_1}{p_0}+(1-p_0)\frac{(1-p_1)}{(1-p_0)}-1\\ &=p_1+(1-p_1)-1\\ &=0\ , \end{aligned} \tag{35}$$

which yields

$$\begin{aligned} \Psi_{d=d_0} &= \frac{1-\left(\frac{\beta}{1-\alpha}\right)}{\left(\frac{1-\beta}{\alpha}\right)-\left(\frac{\beta}{1-\alpha}\right)}\\ &= \frac{1-\alpha-\beta}{1-\alpha}\cdot\frac{\alpha(1-\alpha)}{(1-\alpha)(1-\beta)-\alpha\beta}\\ &= \alpha\ . \end{aligned} \tag{36}$$

In a second step, the expected sample size at termination as a function of the true value d can be approximated by

$$E_d(N) \approx \frac{\Psi_d\log\left(\frac{1-\beta}{\alpha}\right)+(1-\Psi_d)\log\left(\frac{\beta}{1-\alpha}\right)}{p\log\left(\frac{p_1}{p_0}\right)+(1-p)\log\left(\frac{1-p_1}{1-p_0}\right)}\ , \tag{37}$$

where Ψ_d is given by (31).

This dissertation was supported by a grant from the Deutsche Forschungsgemeinschaft (DFG) to the Research Training Group “Statistical Modeling in Psychology” (GRK 2277).

Cover image designed by snowing / Freepik.