Group-Wise Sequential Hypothesis Testing: Approaches and Trade-offs

Shilaan Alzahawi[1]

[1] Stanford University, Graduate School of Business

Author note

All simulation data, analysis code, and research materials have been made publicly available on the OSF and GitHub and can be accessed at osf.io/veczn/ and github.com/shilaan/sequential-testing.

Correspondence concerning this article should be addressed to Shilaan Alzahawi, 655 Knight Way, Graduate School of Business, Stanford, CA 94305. E-mail: shilaan@stanford.edu

Abstract

Sequential analysis allows researchers to perform well-powered experiments with smaller sample sizes. I outline the commonalities and differences between three sequential hypothesis testing procedures: group-sequential designs, the sequential Bayes Factor, and the Independent Segments Procedure. I illustrate how these sequential procedures can be implemented in a group-wise manner, and present two trade-offs of importance to the procedures: long-run error control vs. short-run evidence and efficiency in hypothesis testing vs. accuracy in parameter estimation. Using Monte Carlo simulations, I compare the performance of the procedures in terms of their error control (i.e., false-negative and false-positive decision rates), efficiency (i.e., reduction in required sample size), and accuracy (i.e., bias and variance in effect size estimation). I find that effect size estimates stemming from the Independent Segments Procedure are highly inaccurate, and systematically exclude the true population effect size. I conclude with practical recommendations for researchers who wish to utilize the benefits of sequential hypothesis testing while retaining their ability to obtain a credible effect size estimate.

Keywords: sequential analysis, hypothesis testing, efficiency, estimation, bias

Word count: 11,368

Group-Wise Sequential Hypothesis Testing: Approaches and Trade-offs

In 1929, Harold Dodge and Harry Romig introduced the idea of sequential analysis to the science of statistical quality control (Dodge & Romig, 1929). Sequential analysis is a method of statistical inference that allows data to be analyzed several times during its collection, with the goal of optimizing the amount of resources spent in research (Beffara Bret, Beffara Bret, & Nalborczyk, 2018). When Abraham Wald further developed this idea in the 1940s, his work was deemed so useful to the United States military that it was classified as Restricted under the Espionage Act (Wald, 1973).

Fast-forward 80 years, and we find ourselves in a decade of unparalleled methods reform in psychological science (Nelson, Simmons, & Simonsohn, 2018). At the receiving end of these proposed reforms we often find the statistical power of hypothesis tests (Bakker, van Dijk, & Wicherts, 2012; Button et al., 2013); that is, the long-run probability of observing a statistically significant result ($p < \alpha$) for a given number of subjects and a hypothesized effect size (Morey & Lakens, 2016).

Several suggestions have been made to increase the statistical power of hypothesis tests conducted in psychological science. Not surprisingly, sequential analysis — which allows researchers to perform well-powered experiments at lower cost — has found its way into the limelight (e.g., Beffara Bret et al., 2018; Lakens, 2014; Miller & Ulrich, 2020; Schnuerch & Erdfelder, 2020; Schonbrodt, Zehetleitner, Wagenmakers, & Perugini, 2017).

While research efficiency is commendable, it is not the only advantage conferred by sequential hypothesis testing. Under the Ethics Code of the American Psychological Association, psychological scientists have the ethical obligation to minimize research subjects' exposure to

physical, emotional, or psychological harm (APA, 2019). Sequential analysis can contribute to this goal, by allowing researchers to stop an experiment when the interim data show clear evidence that effects are positive, negative, or largely absent (Proschan, Lan, & Wittes, 2006). During clinical trials, stopping early can be a matter of life or death. Sequential analysis, it turns out, is just as useful today as it was during World War II.

Sequential procedures

In the following section, I briefly outline the sequential procedures under discussion, highlighting several of their commonalities and differences. Next, I introduce the desirable properties of interest to this paper: error control, efficiency, and accuracy. By means of simulation, I will compare the sequential procedures to traditional, fixed-sample hypothesis testing, in terms of false-negative and false-positive error rates, sample size reduction, and the bias and variance involved in effect size estimation.

Group-sequential procedures

The most common procedure for sequential hypothesis testing is a group-sequential (GS) design (Proschan et al., 2006; Schonbrodt et al., 2017). The goal of GS procedures is to efficiently reject a null hypothesis in favor of an alternative hypothesis. In a prototypical group-sequential procedure (e.g., the Pocock correction or the O'Brien-Fleming correction, Fleming, Harrington, & O'Brien, 1984; Pocock, 1977), a researcher is allowed to look at the data several times during the course of a study.

It is well-known that multiple looks at data, conventionally, increase the type I error rate (Simmons, Nelson, & Simonsohn, 2011): i.e., the long-run rate at which a true null hypothesis is rejected. Sequential analysis procedures, however, allow a researcher to take multiple looks at

data without inflating the type I error rate. Critical α levels are adjusted for each look so that the overall type I error rate is controlled at the desired level (for instance, at an α of 5%).

A general, sophisticated approach is provided by an "α-*spending function*" (Lan & Demets, 1983; Wassmer & Brannath, 2016). This approach allows a researcher to have unequally spaced looks at their data (for instance, after the first 50% and the first 75% of the maximum sample size) and to adapt the number of looks as the experiment continues.

For example, a researcher who wishes to control their overall type I error rate at 5% and take k_{max} = 3 looks (i.e., two interim looks and one final look) at the data can use the common Pocock-like α-spending function to determine what the corrected α level should be for each look (referred to as $\alpha_{1,k}$). In this case, $\alpha_{1,1}$ = .023 for the first look, $\alpha_{1,2}$ = .024 for the second look, and $\alpha_{1,3}$ = .033 for the last look (i.e., for a one-tailed test). If during any of the interim looks $p \leq \alpha_1$, the researcher can reject the null hypothesis and end the experiment; this is known as "*stopping for efficacy*."

In addition to stopping early for efficacy, a researcher can "*stop for futility*" if the observed effect size is too small to warrant further investigation. This can be done by manually specifying a binding or non-binding lower bound (for example, a lower bound at $d \leq 0$ or, equivalently, at $p \geq .50$) or by using a "β-*spending function*" to determine the critical lower bound for each look (referred to as $\alpha_{0,k}$). In the remainder of the text, I will use α_1 to refer to the critical value that allows a researcher to stop for efficacy, and α_0 to refer to the critical value that allows a researcher to stop for futility (Wassmer & Brannath, 2016).

The logic of the β-spending function is similar to that of the α-spending function: critical values that warrant stopping for futility are determined a priori while controlling the overall type

II error rate at the desired level (for instance, at a β of 20%). In other words, the bounds are determined such that the probability of a type II error (i.e., given the number of subjects, the long-run probability of failing to reject the null when the true population effect is as hypothesized) is β %.

Figures 1A and 1B provide a visual overview of two possible group-sequential procedures (the Pocock-like and the O'Brien-Fleming-like designs) using α- and β-spending functions, in the case of a one-tailed, two-sample *t*-test with α = 5%, β = 20% for δ = 0.5, and k_{max} = 3 looks. The main difference between the Pocock and the O'Brien-Fleming-like procedure is that the O'Brien-Fleming-like procedure sets a stricter α threshold in earlier looks, with the last look having a threshold closer to the overall α. Consider the α levels for our example Pocock-like design with binding futility bounds: $\alpha_{1,1}$ = .023 for the first look, $\alpha_{1,2}$ = .024 for the second look, and $\alpha_{1,3}$ = .033 for the last look. For the O'Brien-Fleming-like procedure with binding futility bounds, the critical levels are $\alpha_{1,1}$ = .0056 for the first look, $\alpha_{1,2}$ = .022 for the second look, and $\alpha_{1,3}$ = .052 for the last look.

The binding futility bounds for our example Pocock-like design are $\alpha_{0,1}$ = .38 for the first look and $\alpha_{0,2}$ = .13 for the second look. For the O'Brien-Fleming-like design, the binding futility bounds are $\alpha_{0,1}$ = .36 for the first look and $\alpha_{0,2}$ = .14 for the second look. The researcher continues their study until the observed *p*-value crosses one of the prespecified α_1 or α_0 thresholds, or until the maximum *N* has been reached.

Because the OBF-like procedure sets such a strict α threshold for the first look, a researcher employing the O'Brien-Fleming-like procedure will have relatively low power, and consequently a low probability of stopping for efficacy, during the first look. Thus, when a

researcher uses the O'Brien Fleming-like procedure and wishes to be well-powered during the earlier looks, it is often recommended to have the first look take place later in the data collection process (rather than spacing the looks equally; Daniël Lakens, personal communication, December 19, 2020). For example, when $k_{max} = 3$, a researcher can decide to space their looks equally when employing a Pocock-like design (i.e., analyze the data after 33%, 67%, and 100% of the data are collected, respectively), while taking unequally spaced looks when employing an O'Brien-Fleming-like design (e.g., analyze the data after 50%, 75%, and 100% of the data are collected, respectively).

Because the use of group-sequential procedures allows for early stopping, it generally reduces the average sample size needed (with certain exceptions; e.g., when the true effect size is equal to or close to zero, the use of sequential procedures can increase the average sample size needed). For example, if a researcher wants to conduct a one-sided, two-sample *t* test (H_0: $\delta = 0$ against H_1: $\delta > 0$) with 80% power to detect a Cohen's *d* (i.e., a standardized mean difference) of 0.5, a conventional Null Hypothesis Significance Test with an overall α of 5% would require 51 participants per group. Using the Pocock-like GS design outlined above, the same study can be conducted with only 22 participants per group per look. Using the O'Brien-Fleming like design outlined above, this same study can be conducted with 28 participants per group during the first look, and 14 participants per group during the last two looks. Below, I briefly outline the decision-theoretic procedure of GS designs.

Decision-theoretic procedure of GS designs.

1. Randomly sample N_{gs} observations from the specified population(s)
2. Run the test(s) of interest (e.g., a one-tailed, two-sample *t* test) on the obtained sample data

3. During the interim looks (1 to $k_{max} - 1$),
   a. if $p \leq \alpha_1$: decide to reject H_0 and end the experiment ("stop for efficacy")
   b. if $p > \alpha_0$: decide not to reject H_0 and end the experiment ("stop for futility")
   c. if $\alpha_1 < p \leq \alpha_0$: collect N_{gs} additional observations and repeat step 2 on the cumulative data
4. During the final look k_{max} (if reached),
   a. if $p \leq \alpha_1$: decide to reject H_0 and end the experiment
   b. if $p > \alpha_1$: decide not to reject H_0 and end the experiment

Sequential Bayes Factor

Schonbrodt et al. (2017) propose an alternative approach to sequential hypothesis testing: the Sequential Bayes Factor (SBF). With the SBF procedure, decisions are made on the basis of interim Bayes Factors. The Bayes Factor is a measure of relative evidence for two competing hypotheses (e.g., the null and the alternative hypothesis; for an excellent introduction to Bayesian Hypothesis Testing, see Wagenmakers et al., 2018).

The SBF procedure has several properties of interest. First, it is symmetric: it allows researchers to examine whether the obtained data is better predicted by, for example, H_1 than H_0, or vice versa. Second, it is highly flexible: it allows researchers to take as many looks at incoming data as desired (even after every study participant; Schonbrodt et al., 2017). Third, it does not allow type I and type II error rates to be defined a priori (although simulations can be run to choose an optimal design that controls expected error rates, Schonbrodt & Wagenmakers, 2018; Schonbrodt et al., 2017; Stefan, Schonbrodt, Evans, & Wagenmakers, 2020). Fourth, it allows (and requires) researchers to express their uncertainty about the expected effect size by means of a prior distribution.

Again, we consider the example of a researcher running a one-sided, two-sample *t* test between H_0: $\delta = 0$ and H_1: $\delta > 0$. To follow the SBF procedure, the researcher needs to prespecify two things: (a) the prior distribution for the standardized effect size δ under H_1 and (b) the decision threshold *t** for the Bayes Factor. Excellent guidance on these important decisions can be found in Schonbrodt et al. (2017). Ideally, a Bayes Factor Design Analysis or simulations can be run to choose an optimal design that controls expected error rates (Schonbrodt & Wagenmakers, 2018).

To reflect the expectation of a "medium" effect size, a researcher can, for example, choose the default prior distribution (i.e., a Cauchy distribution with scale parameter $r = \sqrt{2}/2$; Morey & Rouder, 2011; Schnuerch & Erdfelder, 2020; Schonbrodt et al., 2017). Alternatively, a researcher can rely on an informed prior distribution that incorporates prior knowledge about the effect size, gathered from expert knowledge, previous work, or theoretical considerations (Stefan et al., 2020). In addition, the researcher could (for example, on the basis of simulations) choose BF_{10} = 3 as a decision threshold (commonly referred to as "*moderate evidence*"; Schonbrodt et al., 2017). This decision threshold *t** implies that the researcher is content to stop sampling once the data is 3 times more likely under H_1 than under H_0, or vice versa. Figure 2A (2B) provides a visual example of a situation in which the Bayes Factor indicates that the data is 3 times more likely under H_1 (H_0).

The Sequential Bayes Factor and the Group-Sequential Procedures address different questions, each interesting in their own right. Group-sequential procedures (as well as the traditional Null Hypothesis Significance Test and the Independent Segments Procedure described below) raise the question "*On the basis of a decision rule under which the long-run probability of rejecting a true null hypothesis will be no more than α%, can the null hypothesis be rejected?*"

The Sequential Bayes Factor procedure, on the other hand, considers "*On the basis of a prespecified threshold and desired error rates under the assumption of a specific H_0 and H_1, can we conclude that there is sufficient evidence for H_0 or H_1?*"

There are several possible ways to implement the SBF procedure (Schonbrodt & Wagenmakers, 2018). An open-ended SBF design is one in which a researcher tests after each participant and stops data collection when sufficient evidence for either H_1 or H_0 has been gathered. Alternatively, a researcher can prespecify a maximum sample size and continue data collection until (a) sufficient evidence has been gathered or (b) the maximum sample size has been reached. In this paper, I focus on the latter procedure: a max-SBF design. Below, I briefly outline the decision-theoretic procedure of the max-SBF.

Decision-theoretic procedure of the Sequential Bayes Factor with maximum N.

1. Randomly sample N_{Bayes} observations from the specified population(s)
2. Run the test(s) of interest (e.g., a one-sided default Bayes Factor two-sample *t* test)
   a. if $BF_{10} \geq t*$: decide to reject H_0 (support H_1) and end the experiment
   b. if $BF_{10} \leq 1/t*$: decide to reject H_1 (support H_0) and end the experiment
   c. if $1/t* < BF_{10} < t*$: collect additional observations and repeat step 2 on the cumulative data
3. Continue until the Bayes Factor crosses one of the thresholds, or until the maximum *N* has been reached

Independent segments procedure

Another sequential procedure recently proposed for use in psychological research is the Independent Segments Procedure (ISP; Miller & Ulrich, 2020). A central property of the ISP is that data is collected in a series of independent experiments ('segments'). Using other sequential

procedures (such as the GS designs and the Sequential Bayes Factor), data collection and analysis is cumulative: when an experiment is continued (because the interim data did not cross the pre-specified boundaries for early stopping), a researcher collects more observations, adds these to the observations from previous looks, and calculates an interim test statistic on the basis of all the data collected so far. In contrast, when an independent segments experiment is continued, the researcher discards the previous batch of observations and calculates an interim test statistic on the basis of a new, completely independent, batch of observations. Unlike the SBF and GS-designs, the ISP ignores data from earlier looks. The ISP, by design, incorporates information loss.

In addition, the ISP does not have the same flexibility as the procedures described above, in two important regards. First, unlike the alternative research strategies discussed in this paper, the ISP requires that a researcher sets the exact number of participants and segments in advance, and all segments are assumed to be equally spaced (i.e., when k_{max} = 2 segments, the first look at the data occurs when exactly 50% of the maximum sample size has been collected). Thus, the ISP will be less appropriate for researchers with practical constraints who, for example, cannot guarantee the exact same sample size in each respective segment. Again, consider our example of a one-tailed, two-sample *t* test with k_{max} = 3 segments. Figures 3A and 3B provide a visual overview of the independent segments procedure, compared to the conventional fixed-sample Neyman-Pearson hypothesis test (Figure 3A shows the conceptual differences; Figure 3B shows our worked example). In order to achieve 80% power with an overall α of 5% and a hypothesized population effect size δ = 0.5, the required sample size N_{ISP} = 25 per group per segment. If the researcher deviates in any way from having exactly 25 participants per group per segment, inferences on the basis of the ISP will be affected. Although this influence may generally be

small for practical purposes, there are cases in which the effects of violating an assumption of equally sized stages are non-negligible (Wassmer & Brannath, 2016).

Second, the ISP only applies to research contexts in which a single, key hypothesis is tested (Miller & Ulrich, 2020). This, unfortunately, is unusually restrictive to the settings in which sequential hypothesis testing is common. For example, medical trials often have multiple "*endpoints*" that need to be monitored at the same time: a primary endpoint (e.g., treatment success or survival) and multiple secondary – but exceedingly important – efficacy or safety endpoints (Wassmer & Brannath, 2016).

In addition, unlike the Sequential Bayes Factor and the group-sequential designs with futility bounds, the ISP does not incorporate a principled decision rule to reject the alternative hypothesis and/or conclude support for the null hypothesis (Lakens, 2021a). The Bayes Factor is an inherently comparative measure of evidence, and can be used to accept the null hypothesis based on the crossing of a prespecified threshold. It is well known that the threshold for concluding evidence of absence on the basis of a Bayes Factor is quite stringent: concluding support for the null demands a relatively large number of observations (Stefan, Gronau, Schonbrodt, & Wagenmakers, 2019). Group-sequential designs also allow researchers to reject the alternative hypothesis in a principled manner, by setting futility bounds on the basis of a β-spending function that controls the overall type II error rate. Unlike the SBF and group-sequential procedures, researchers relying on the ISP cannot reject the alternative hypothesis in a manner that either demands crossing a (rather stringent) threshold for absence of evidence or controls overall type II error rates. Indeed, Miller and Ulrich (2020, p. 6) admit that the Independent Segments Procedure "would not be useful for studies designed to produce evidence for the absence of an effect."

In sum, the ISP applies only to those settings in which a researcher is comfortable with discarding inconclusive data; can guarantee an exact number of participants, allocated equally across segments; wishes to test a single, key hypothesis; and does not desire a principled decision rule to reject the alternative hypothesis. Below, I briefly outline the decision-theoretic procedure of the ISP.

Decision-theoretic procedure of the ISP.

1. Randomly sample N_{ISP} observations from the specified population(s)
2. Run the test(s) of interest (e.g., a one-tailed, two-sample *t* test) on the obtained sample data
3. For segments 1:$k_{max} - 1$,
   a. if $p \leq \alpha_{strong}$: decide to reject H_0 and end the experiment
   b. if $p > \alpha_{weak}$: decide not to reject H_0 and end the experiment
   c. if $\alpha_{strong} < p \leq \alpha_{weak}$: discard the data, collect a new batch of N_{ISP} observations, and repeat step 2 on this independent batch of data
4. For segment k_{max} (if reached), repeat steps 1 and 2,
   a. if $p \leq \alpha_{weak}$: decide to reject H_0 and end the experiment
   b. if $p > \alpha_{weak}$: decide not to reject H_0 and end the experiment

Table 1 summarizes a selection of commonalities and differences between the sequential hypothesis testing procedures under discussion.

Common trade-offs in the scientific research process

The scientific research process often involves mutually incompatible goals (McGrath, 1981). At times, scientific goals that are each worthy in their own right can conflict: optimizing for one goal tends to decrease our ability to achieve the other (e.g., Goodman, 2007). In the following section, I introduce two such pairs of scientific goals. First, I elaborate on the trade-off between

controlling errors in the long-run versus obtaining compelling evidence for the case at hand. Second, I describe the trade-off between efficiently testing hypotheses versus accurately estimating effect sizes. In addition, I explain why these trade-offs (long-run error control versus short-run evidence and efficiency versus accuracy, respectively) are pertinent to our discussion of sequential hypothesis testing.

Long-run error control versus short-run evidence

"*Today, I speak to you of war. A war that has pitted statistician against statistician for nearly one hundred years. A mathematical conflict that has recently come to the attention of the normal people. And these normal people look on in fear, in horror, but mostly in confusion because they have no idea why we're fighting*" (Lawrence Livermore National Laboratory, 2016). Such was the introduction of statistician Kristin Lennox to a talk at the Lawrence Livermore National Laboratory. The war of which she speaks is the war between Bayesian and frequentist statistics.

In 1933, Jerzy Neyman and Egon Pearson introduced an ingenious device for testing competing hypotheses (Neyman & Pearson, 1933). The goal of Neyman-Pearson hypothesis testing is to decide between two competing hypotheses with explicit control over two types of error, introduced earlier in this paper: type I error (the long-run rate at which a true null hypothesis is rejected) and type II error (for a given nonzero population effect size, the long-run rate at which the null hypothesis is maintained). By definition, this is a frequentist procedure: the goal is to devise a decision rule for choosing between competing hypotheses under which the decision maker, in the long run (i.e., when considering the frequency of all possible outcomes), will not be wrong too often (Goodman, 1999; Neyman & Pearson, 1933).

This ingenious device, however, comes at a price that its users should be willing to pay. The outcome of the Neyman-Pearson hypothesis test is a behavior—a decision to reject or accept a hypothesis. This behavior has no short-run (i.e., experiment-specific) meaning: it is not an inference about the relative evidence for competing hypotheses provided by the experiment at hand. According to the Neyman-Pearson standpoint, it is impossible to provide any (relative) evidence of the truth or falsity of the hypotheses under study (Neyman & Pearson, 1933). All we can say is that a decision was made on the basis of a rule that, in the long-run, controls error probabilities (Dienes, 2011). To enjoy the benefits of error control provided by Neyman-Pearson hypothesis testing, "[w]e must abandon our ability to measure evidence, or judge truth, in an individual experiment" (Goodman, 1999: p. 998).

It is at this junction — between the ability to measure evidence in a single experiment and the ability to control the number of mistaken conclusions in the long-run — that we meet the Bayesian statistician. According to the Bayesian, evidence is a property of the obtained data that makes us alter our belief about the hypotheses at hand. The Bayesian measure of evidence, previously encountered by the reader, is the Bayes Factor. There is, however, no free lunch in statistical inference (Rouder, Morey, Verhagen, Province, & Wagenmakers, 2016). Because the Bayesian calculation considers no outcomes other than the one observed, robust error control can no longer be ensured (Dienes, 2011; Kruschke & Liddell, 2018). The consideration of outcomes that *could* be observed (i.e., the tail area of the sampling distribution) under a hypothesis of interest is essential to ensuring a test's long-run performance and associated error probabilities (Mayo, 2018).

To some, the use of tail areas in hypothesis testing is an admirable feature that ensures long-run error control; to others, it is a bug that defies our ability to articulate the evidence for

competing hypotheses provided by the experiment at hand. Controlling (long-run) error rates and quantifying (short-run) evidential strength are both worthy scientific goals. However, you can't have one without implicitly abstaining from the other (unless you follow the Likelihood paradigm of statistical inference, which is beyond the scope of this paper; Blume, 2002; Goodman & Royall, 1988; Schnuerch & Erdfelder, 2020). Ultimately, the question is what is more important to the researcher (Lakens, 2021b): controlling long-run error rates or quantifying short-run evidential strength (Dienes, 2011).

Why does this matter to our discussion of sequential analysis? It may seem that we have strayed far from the comparison of sequential hypothesis testing procedures. However, we are right at its core. As follows from the above discussion, the Sequential Bayes Factor – unlike a GS design or the ISP – is not guaranteed to control type I and type II error probabilities. Thus, it is interesting to evaluate the practical implications of the philosophical differences between frequentist and Bayesian hypothesis testing. I will return to this issue in the following section of the paper, in which I compare the error rates of the Sequential Bayes Factor to the fixed-sample Neyman-Pearson hypothesis test, the Pocock-like and O'Brien-Fleming like group-sequential procedures, and the Independent Segments Procedure.

Efficiency versus accuracy

Of special interest to our discussion of sequential analysis, a trade-off exists between efficiently testing hypotheses and accurately estimating an effect size (Schonbrodt et al., 2017). Accurate parameter estimation often requires more observations than a (sequential) hypothesis tester is willing or able to collect. Efficiency and accuracy are each worthy scientific goals in their own right and, unfortunately, often conflict (Goodman, 2007).

The goal of sequential procedures is to efficiently test hypotheses; the goal is not to properly estimate effect sizes (Schonbrodt et al., 2017). It is well known that early stopping can seriously inflate effect sizes (Fan, DeMets, & Lan, 2004; Miller & Ulrich, 2020; Pocock & Hughes, 1989). However, when taking all studies into account (including those that did not terminate early), the bias inherent in both group-sequential designs (Fan et al., 2004; Goodman, 2007) and Sequential Bayes Factors (Schonbrodt et al., 2017) tends to be small. Thus, group-sequential designs and Sequential Bayes Factors can increase research efficiency at only a slight cost to our ability to estimate the true effect size. No such investigation, however, has been conducted for the Independent Segments Procedure.

Of note, because the properties of group-sequential designs are well-studied and understood, bias correction procedures are now widely available for group-sequential designs. That is, rather than relying on a naive, uncorrected effect size estimate stemming from a sequential design (e.g., Cohen's *d*), researchers can use a median unbiased estimator (which can be easily implemented in widely used sequential analysis software such as `rpact`; Wassmer & Pahlke, 2021). A similar bias-correction procedure can be followed for the Sequential Bayes Factor: instead of relying on a naive, uncorrected effect size estimate, a researcher employing a SBF design can use the mean of the posterior distribution as the effect size estimator, which shrinks the estimate in early terminations and thus counteracts the effect size inflation caused by early stopping (Schonbrodt et al., 2017). No such bias correction procedure, however, exists for the Independent Segments Procedure. Thus, it is of great interest to study the accuracy of effect size estimation when relying on the Independent Segments Procedure.

In the following section, I use Monte Carlo simulations to compare group-sequential designs, Sequential Bayes Factors, and the Independent Segments Procedure to the fixed-sample

Neyman-Pearson procedure, focusing on their relative error rates, efficiency, and accuracy. In a previous contribution, the Sequential Bayes Factor was compared to GS designs in terms of error rates, efficiency, and accuracy (Schonbrodt et al., 2017). However, these comparisons were unbalanced in two regards: (a) the GS designs used did not allow for early stopping (i.e., they did not incorporate, for instance, a β-spending function to stop for futility) and (b) the number of looks was not equal across procedures (i.e., the Sequential Bayes Factor procedure was based on a near-unlimited number of looks until the decision threshold was met, while the chosen GS design incorporated only four looks).

Schonbrodt et al. (2017) show that a highly flexible implementation of the Sequential Bayes Factor – i.e., the SBF in all its glory, with as many looks as is desired – is at least as efficient as a default group-sequential design (i.e., a group-sequential approach that does not incorporate binding futility bounds on the basis of a β-spending approach). In this paper, I offer a more balanced comparison between the two procedures. First, because the group-sequential approach also offers a principled decision rule to reject an alternative hypothesis, I implement the group-sequential designs with binding futility bounds. Second, I implement the SBF with k_{max} equally spaced looks, for two reasons: (a) to allow for maximum comparability between the sequential procedures under discussion, and (b) to account for practical considerations; it may very well be impractical for researchers to implement the maximum-flexibility SBF approach and analyze data after every participant. Thus, I focus on a slightly different, pragmatic question: For a fixed number of looks at data (i.e., a "group-wise" implementation), how do these respective procedures perform in terms of their error rates, efficiency, and accuracy?

Method

By means of simulation, I will demonstrate the properties of the procedures under discussion in the context of testing hypotheses about positive mean differences between two independent groups (i.e., a one-tailed, two-sample *t*-test). The procedures will be compared in terms of three main properties: (a) error rates, (b) efficiency, and (c) accuracy in effect size estimation. We will compare the properties of the three sequential procedures to a fixed-sample hypothesis test (i.e., a conventional Null Hypothesis Significance Test) paired with an equivalence test.

For maximum comparability to the sequential procedures that allow some principled form of support for the null hypothesis (i.e., the Sequential Bayes Factor and the group-sequential designs with binding futility bounds), we combine the fixed-sample hypothesis test with an equivalence test. Conventionally, the Null Hypothesis Significance test is not able to distinguish between failures to reject the null hypothesis due to evidence of absence or absence of evidence. Pairing the fixed sample hypothesis test with an equivalence test overcomes this limitation, by allowing us to reject the presence of a difference more extreme than a smallest effect size of interest (Lakens, Scheel, & Isager, 2018). In the following section, I briefly outline how each property under study – error rates, efficiency, and accuracy in effect size estimation – is evaluated.

Properties of Interest: Error Rates, Efficiency, and Accuracy

Decision-theoretic error rates are evaluated by classifying test results as True Positive, True Negative, False Positive, False Negative, or Inconclusive. Below, I outline the cases in which the outcomes of a hypothesis test were labelled as True Positive, True Negative, False Positive, False Negative, or Inconclusive, respectively.

True Positive.

1. For the fixed sample hypothesis test paired with an equivalence test, $\delta > 0$ and the decision was made to reject the null hypothesis ($p \leq \alpha$)
2. For the Pocock and O'Brien-Fleming procedures, $\delta > 0$ and the decision was made to reject the null hypothesis ($p \leq \alpha_{1,k}$)
3. For the SBF procedure, $\delta > 0$ and the decision was made to reject the null hypothesis ($BF_{10} \geq t^*$)
4. For the Independent Segments Procedure, $\delta > 0$ and the decision was made to reject the null hypothesis ($p \leq \alpha_{strong}$ in segments $1{:}k_{max} - 1$ or $p \leq \alpha_{weak}$ in the final segment k_{max})

False Positive.

1. For the fixed-sample hypothesis test paired with an equivalence test, $\delta = 0$ and the decision was made to reject the null hypothesis ($p \leq \alpha$)
2. For the Pocock and O'Brien-Fleming procedures, $\delta = 0$ and the decision was made to reject the null hypothesis ($p \leq \alpha_{1,k}$)
3. For the SBF procedure, $\delta = 0$ and the decision was made to reject the null hypothesis ($BF_{10} \geq t^*$)
4. For the Independent Segments Procedure, $\delta = 0$ and the decision was made to reject the null hypothesis ($p \leq \alpha_{strong}$ in segments $1{:}k_{max} - 1$ or $p \leq \alpha_{weak}$ in the final segment k_{max})

True Negative.

1. For the fixed-sample hypothesis test paired with an equivalence test, $\delta = 0$ and the decision was made to conclude statistical equivalence ($p \leq \alpha$ for both *p*-values of a Two One-Sided Tests [TOST] procedure; D. Lakens et al., 2018)
2. For the Pocock and O'Brien-Fleming procedures, $\delta = 0$ and the decision was made to stop for futility ($p \geq \alpha_{0,k}$)
3. For the SBF procedure, $\delta = 0$ and the decision was made to support the null hypothesis ($BF_{01} \geq t^*$)
4. For the Independent Segments Procedure, $\delta = 0$ and the decision was made to stop in an early segment ($p > \alpha_{weak}$)

False Negative.

1. For the fixed-sample hypothesis test paired with an equivalence test, $\delta > 0$ and the decision was made to conclude statistical equivalence ($p_{\text{TOST},1:2} \leq \alpha$)
2. For the Pocock and O'Brien-Fleming procedures, $\delta > 0$ and the decision was made to stop for futility ($p \geq \alpha_{0,k}$)
3. For the SBF procedure, $\delta > 0$ and the decision was made to support the null hypothesis ($BF_{01} \geq t^*$)
4. For the Independent Segments Procedure, $\delta > 0$ and the decision was made to stop in an early segment ($p > \alpha_{weak}$)

Inconclusive.

1. For the fixed-sample hypothesis test paired with an equivalence test, the procedure failed to reject the null hypothesis or conclude statistical equivalence ($p_{\text{fixed}} > \alpha$ and $p_{\text{TOST},1:2} > \alpha$)
2. For the Pocock and O'Brien-Fleming procedures, k_{max} was reached and ended in a failure to reject the null hypothesis ($p > \alpha_1$)
3. For the SBF procedure, k_{max} was reached and neither of the thresholds was met ($1/t* < BF_{10} < t*$)
4. For the Independent Segments Procedure, k_{max} was reached and ended in a failure to reject the null hypothesis ($p > \alpha_{weak}$)

Procedural efficiency is evaluated by means of the average (expected) sample size across several true effect sizes for each procedure. Accuracy in effect size estimation is evaluated by means of the density of empirical effect size estimates, the bias in effect size estimates (i.e., the median estimated effect size compared to the true effect size), and the mean squared error of effect size estimates (i.e., the sum of the squared bias and the variance of effect size estimates) derived from each procedure. For all procedures, accuracy is evaluated for the obtained, uncorrected effect size estimates. In addition, for procedures that have a bias-adjusted estimator available (i.e., the

Sequential Bayes Factor and the group-sequential designs), I also evaluate the accuracy of the bias-corrected estimates.

Settings of the Simulation

I simulate populations with a specific standardized mean difference δ and examine the efficiency (i.e., average sample size required), error rates (i.e., rates of false-positive and false-negative evidence), and accuracy (i.e., of the simulation estimate of the population effect size) of five hypothesis testing procedures: the fixed-sample Neyman-Pearson hypothesis test paired with an equivalence test (Lakens, 2014), a group-sequential design with Pocock-like α- and β-spending functions (Pocock, 1977; Wassmer & Brannath, 2016), a group-sequential design with O'Brien-Fleming-like α- and β-spending functions (Fleming et al., 1984; Wassmer & Brannath, 2016), the sequential Bayes Factor (Schonbrodt et al., 2017), and the Independent Segments Procedure (Miller & Ulrich, 2020).

To facilitate discussion, I focus on one typical scenario in reporting the simulation results. For the frequentist procedures (i.e., the fixed-sample N-P procedure paired with an equivalence test, the ISP, and the Pocock-like and O'Brien-Fleming-like GS designs, respectively) under discussion, the chosen typical scenario corresponds to a one-tailed, two-sample hypothesis test powered to 80% (β = 0.2) to detect a hypothesized population effect size of δ = 0.5 with α = .05. In the fixed-sample case, these specifications require a sample size of 51 observations per group. The equivalence bounds to achieve 80% power with N = 51 are a lower equivalence bound of Δ_L = -0.58 and an upper equivalence bound of Δ_U = 0.58.

All sequential procedures, with the exception of the O'Brien-Fleming-like GS design, are set to a maximum number of 3 equally spaced looks (i.e., the interim data are looked at twice,

after 33% and 67% of the data are collected, and the final look takes place when all data are collected). In contrast, because the O'Brien-Fleming-like GS design sets such a stringent α threshold in the first look, it is often recommended to have the first look take place on a larger fraction of the data (Daniël Lakens, personal communication, December 19, 2020). In this case, I implemented the O'Brien-Fleming-like GS design with the two interim looks taking place after 50% and 75% of the data, respectively, are collected.

For the independent segments procedure with α_{total} = .05, α_{strong} = .025, α_{weak} = .28, β_{total} = .2 and k_{max} = 3 segments, the required sample size N_{ISP} = 25 per group per segment. For the GS designs with α_{total} = .05, β_{total} = .20, k_{max} = 3 looks, an α-spending approach was used to determine critical $\alpha_{1,k}$ levels. For the example discussed here, the adjusted $\alpha_{1,k}$ levels are $\alpha_{1,1}$ = .023, $\alpha_{1,2}$ = .024, and $\alpha_{1,3}$ = .033 at the first, second, and third look for the Pocock-like design and $\alpha_{1,1}$ = .0056, $\alpha_{1,2}$ = .022, and $\alpha_{1,3}$ = .052 at the first, second, and third look for the O'Brien-Fleming-like design.

In addition, a β-spending function was used to set critical $\alpha_{0,k}$ levels (i.e., binding futility bounds). For our worked example, the adjusted $\alpha_{0,k}$ levels correspond to $\alpha_{0,1}$ = .38 and $\alpha_{0,2}$ = .13 at the first and second look for the Pocock-like design and $\alpha_{0,1}$ = .36 and $\alpha_{0,2}$ = .14 at the first and second look for the O'Brien-Fleming-like design. For the Pocock-like design, the required sample size N_{Pocock} = 22 per group per look; for the O'Brien-Fleming-like design, the required sample size $N_{\text{O'Brien-Fleming}}$ = 28 per group for the first look, and 14 per group for the last two looks.

The parameters for the Sequential Bayes Factor procedure were chosen to allow for maximum comparability with the frequentist sequential procedures, which correspond to a one-

tailed hypothesis test powered to detect a hypothesized population effect size of δ = 0.5 with α = .05 and $k_{max} = 3$ looks. To calculate the Bayes Factor, I used an informed prior (a normal distribution centered at the hypothesized effect size, $\delta = 0.5$). To incorporate information about the sidedness of the hypothesized effect, the prior distribution was truncated at $\delta = 0$ to include only positive values. Thus, like in the one-sided *t*-tests described above, we compare H_0: $\delta = 0$ to H_1: $\delta > 0$.

For the SBF procedure, I ran simulations to choose an optimal sample size (paired to the other procedures under study; a group-wise implementation with k_{max} = 3 equally spaced looks) and Bayes Factor threshold, such that the overall error rate would approach that of the frequentist procedures (i.e., a false negative rate of 20% and a false positive rate of 5%). In the main paper, I report the results of a max-SBF procedure with the Bayes Factor threshold set to 3 for H_1 and 1/3 for H_0, the sample size set to N_{Bayes} = 25 per group per look with k_{max} = 3 looks, and a normal prior distribution centered at $\delta = 0.5$ with $\sigma^2 = 0.3$. Of note, as discussed above, the Sequential Bayes Factor can be implemented much more flexibly (i.e., with as many looks as desired, without a maximum sample size). However, to make the sequential procedures as comparable as possible and to account for practical considerations that researchers may have, each procedure is each implemented in a group-wise manner, with k_{max} = 3 looks.

For the simulations, I drew random samples from two normal distributions with common variance σ^2 = 1 and means μ_1 = 0 and $\mu_2 = \delta$ (δ = -0.2, 0, 0.2, 0.4, 0.5, 0.6, 0.8, 1). For each hypothesis testing procedure and true effect size δ, 10,000 iterations were simulated (i.e., 5 procedures with 8 distinct true effect sizes and 10,000 replications each, leading to 400,000 observations in total). Below, I briefly outline each of the five decision-theoretic hypothesis testing procedures followed.

Fixed-sample hypothesis test

1. Randomly sample N_{fixed} observations from each of the two specified populations
2. Run a one-tailed, two-sample *t* test on the mean difference between the two samples
    a. If $p \leq \alpha$: decide to reject H_0
    b. If $p > \alpha$: decide not to reject H_0
3. Run a one-tailed, two-sample equivalence test (TOST) on the mean difference between the two samples
    a. If $p_{\text{TOST,1:2}} \leq \alpha$: decide to reject H_1
    b. If $p_{\text{TOST,1:2}} > \alpha$: decide not to reject H_1 (inconclusive)
4. To estimate the standardized effect size, calculate Cohen's *d* (the difference in sample means divided by their pooled standard deviation)
5. Repeat 10,000 times

Pocock-like (O'Brien-Fleming-like) Group-sequential procedures

1. Randomly sample N_{Pocock} ($N_{\text{O'Brien-Fleming}}$) observations from each of the two specified populations
2. Run a one-tailed, two-sample *t* test on the mean difference between the two samples
3. For segments 1:$k_{max} - 1$,
    a. if $p \leq \alpha_1$: decide to reject H_0 and end the experiment ("stop for efficacy")
    b. if $p > \alpha_0$: decide to reject H_1 and end the experiment ("stop for futility")
    c. if $\alpha_1 < p \leq \alpha_0$: collect N_{Pocock} ($N_{\text{O'Brien-Fleming}}$) additional observations and repeat step 2 on the cumulative data
4. For segment k_{max} (if reached),
    a. if $p \leq \alpha_1$: decide to reject H_0 and end the experiment
    b. if $p > \alpha_1$: decide not to reject H_0 and end the experiment (inconclusive)
5. To estimate the naive (i.e., uncorrected) standardized effect size, calculate Cohen's *d*
6. To estimate the bias-adjusted standardized effect size, calculate the median unbiased effect estimate
7. Repeat 10,000 times

Sequential-Bayes Factor

1. Randomly sample N_{Bayes} observations from each of the two specified populations
2. Run a one-sided, two-sample Bayes Factor *t* test on the mean difference between the two samples
    a. if $BF_{10} \geq 3$: decide to reject H_0 (support H_1) and end the experiment
    b. if $BF_{10} \leq 1/3$: decide to reject H_1 (support H_0) and end the experiment
    c. if $1/3 < BF_{10} < 3$: collect N_{Bayes} additional observations and repeat step 2 on the cumulative data
3. Continue until the Bayes Factor crosses one of the thresholds, or until the maximum *N* ($N_{\text{Bayes}} * k_{max}$) has been reached
4. To estimate the uncorrected standardized effect size, calculate Cohen's *d*
5. To estimate the bias-adjusted standardized effect size, calculate the mean of the posterior distribution using a two-sided model (Van Doorn et al., 2020)
6. Repeat 10,000 times

Independent segments procedure

1. Randomly sample N_{ISP} observations from each of the two specified populations
2. Run a one-tailed, two-sample *t* test on the mean difference between the two samples
3. For segments 1:$k_{max} - 1$,
    a. if $p \leq \alpha_{strong}$: decide to reject H_0 and end the experiment
    b. if $p > \alpha_{weak}$: decide not to reject H_0 and end the experiment
    c. if $\alpha_{strong} < p \leq \alpha_{weak}$: discard the data, collect N_{ISP} new observations, and repeat step 2 on this independent batch of data
4. For segment k_{max} (if reached), repeat steps 1 and 2,
    a. if $p \leq \alpha_{weak}$: decide to reject H_0 and end the experiment
    b. if $p > \alpha_{weak}$: decide not to reject H_0 and end the experiment
5. To estimate the uncorrected standardized effect size, calculate Cohen's *d*
6. Repeat 10,000 times

Transparency and openness

All data, code, and research materials are available at osf.io/veczn and github.com/shilaan/sequential-testing. The study was not preregistered. This manuscript was created using R (Version 4.1.2; R Core Team, 2021) and the R-packages *BayesFactor* (Version 0.9.12.4.2; Morey & Rouder, 2018), *BFDA* (Version 0.5.0; Schönbrodt & Wagenmakers, 2018), *citr* (Version 0.3.2; Aust, 2019), *coda* (Version 0.19.4; Plummer, Best, Cowles, & Vines, 2006), *doParallel* (Version 1.0.16; Corporation & Weston, 2020), *dplyr* (Version 1.0.7; Wickham, François, Henry, & Müller, 2021), *forcats* (Version 0.5.1; Wickham, 2021a), *foreach* (Version 1.5.1; Microsoft & Weston, 2020), *ggplot2* (Version 3.3.5; Wickham, 2016), *gridExtra* (Version 2.3; Auguie, 2017), *iterators* (Version 1.0.13; Analytics & Weston, 2020), *Matrix* (Version 1.3.4; Bates & Maechler, 2021), *papaja* (Version 0.1.0.9997; Aust & Barth, 2020), *pracma* (Version 2.3.3; Borchers, 2021), *purrr* (Version 0.3.4; Henry & Wickham, 2020), *readr* (Version 2.0.2; Wickham & Hester, 2021), *rpact* (Version 3.1.1; Wassmer & Pahlke, 2021), *stringr* (Version 1.4.0; Wickham, 2019), *tibble* (Version 3.1.5; Müller & Wickham, 2021), *tidyr* (Version 1.1.4; Wickham, 2021b), *tidyverse* (Version 1.3.1; Wickham et al., 2019), *tinylabels* (Version 0.2.1; Barth, 2021), and *TOSTER* (Version 0.3.4; Lakens, 2017).

Results

Error rates

The error rates involved in the procedures under discussion are shown in Table 2A, Table 2B, and Figure 4. Table 2A describes the rates of True Negative, Inconclusive, and False Positive results when the true population effect size $\delta = 0$; Table 2B describes the rates of True Positive, Inconclusive, and False Negative results when the true population effect size $\delta = 0.50$. Figure 4

shows the True Positive, Inconclusive, and False Negative rates for all population effect sizes greater than zero (δ = 0.2, 0.4, 0.6, 0.8, 1). As can be read from the tables, the frequentist procedures all perform as expected: 5% of simulations rejected the null hypothesis when δ = 0 (i.e., a type I error rate of 5%; see the False Positive column Table 2A) and 80-82% of simulations rejected the null hypothesis when δ = 0.5 (i.e., a type II error rate of 18-20%; see the True Positive column of Table 2B).

In addition, the results speak to the ability of using simulations to design an efficient Bayesian hypothesis testing procedure with attractive long-run properties. In this case, the False Positive rate of the SBF procedure is 6% and the True Positive rate of the SBF procedure is 84%. In line with previous findings (Schonbrodt & Wagenmakers, 2018; Schonbrodt et al., 2017; A. Stefan et al., 2020), I find that simulation-based approaches allow researchers to choose an optimal Sequential Bayes Factor design that controls expected error rates. It has previously been shown that the maximum-flexibility SBF procedure – with appropriate choices for the decision threshold and the prior distribution of effect sizes under the alternative hypothesis – can perform (at least) as well as other (sequential) procedures in terms of false-positive and false-negative error rates (Schonbrodt et al., 2017; A. Stefan et al., 2020). Here, I show that this observation holds when we use a restricted version of the SBF with k_{max} = 3 equally spaced looks. Continuous monitoring is often impractical (Jennison & Turnbull, 2000). Thus, it is of great interest that the SBF may be implemented in a group-wise manner, with a guaranteed upper limit on sample size, while retaining the ability to control expected error rates.

Efficiency

Table 3 and Figure 5 show the average sample size required for the procedures under investigation, as a function of the true population effect size. As discussed, the fixed-sample

procedure requires 51 participants per group. As can be read from Table 3 and Figure 5, the sequential procedures allow for a reduction in the average sample size required. Across the procedures, efficiency gains appear to be rather similar. When the true population effect size is relatively small, the max-SBF procedure is less efficient than its frequentist counterparts. When the true population effect size is relatively large, the O'Brien-Fleming approach is slightly less efficient than its counterparts. This can be explained by the relatively stringent α_1 levels of the O'Brien-Fleming procedure in the first segment (recall that, for the O'Brien-Fleming-like procedure, $\alpha_{1,1}$ = .0056). Across all true effect sizes, the Pocock-like procedure, with our chosen parameters, provides the greatest efficiency gains.

Our sequential procedures are performing as expected, providing an appealing increase in the efficiency of hypothesis tests. Next, we consider whether — and if so, to what extent — these efficiency gains come at a cost to our ability to estimate the true effect size.

Accuracy

Figure 6 shows the empirical distribution of effect size estimates, along with the available bias-adjusted estimates, stemming from the five hypothesis testing procedures when the true effect size δ = 0.5. Effect size estimates stemming from the fixed sample hypothesis test are, as expected, normally distributed around the true effect size (δ = 0.5). As we can see, the distribution of effect size estimates stemming from the sequential procedures have a somewhat odd shape. This is to be expected (see, for example, Goodman, 2007). The distribution of observed effects is more spread out for the sequential procedures because the experiments that stopped early have smaller sample sizes and higher variability in effect size estimates. In addition, we observe left-tailed peaks in the distribution of observed effect sizes. This is a

consequence of those experiments that stopped early with a decision to support the null (i.e., False Negative results).

For the group-sequential designs, we can see that the available median-unbiased estimators provide a great improvement over the naive (Cohen's *d*) estimator. For both the Pocock-like and the O'Brien-Fleming-like designs, the bias-adjusted effect size estimate is very close to the true population effect size: $d = 0.49$ compared to $\delta = 0.50$. For the SBF procedure, we see that the using the mean of the posterior distribution (which shrinks the effect size estimate in earlier terminations) also provides a slight improvement in parameter estimation (by reducing the estimated effect size from $d = 0.57$ to $d = 0.55$). The bias in the median effect size estimate of the ISP, however, is severe ($d = 0.66$; an exaggeration ratio of 1.33). This biased estimate remains uncorrected due to the absence of a bias-adjusted estimator accompanying the ISP. A we can see, the distribution of effect size estimates stemming from the ISP is bimodal.

To make sense of this result, we revisit Figure 3B. In the first and second segment of the ISP, the critical effect size (i.e., the effect size that needs to be observed to reject the null hypothesis) is $d = 0.57$. An effect $d \geq 0.57$ leads to an early stop with a decision to reject the null hypothesis; an effect $d \leq 0.16$ leads to an early stop with a decision not to reject the null hypothesis. An observed effect size $0.16 \leq d < 0.57$ is considered inconclusive, after which the data is discarded and a new batch of observations is collected. As a result, *effect sizes closest to the true population effect size are discarded*.

In figure A1 in Appendix A, the reader can find that this observation is not merely idiosyncratic to the test chosen (a two-sample *t* test with $\alpha = .05$ and $\beta = .20$ for a hypothesized population effect size $\delta = 0.5$). In contrast, across dozens of different test specifications, effect size estimates stemming from the ISP are consistently bimodal and systematically exclude the

true population effect size from being observed. Because the ISP involves running three independent mini-experiments (rather than accumulating data over time), the individual segments – with the exception of the last segment, for which a very weak α is set – are severely underpowered. The first $k_{max} - 1$ segments of the ISP tend to have less than 50% power. As a result, the critical effect size for these segments will always be higher than the hypothesized effect size. In our worked example, the first two individual segments have a power of 41%. To make sure that the hypothesized effect size is not consistently weeded out by the procedure's significance filter, the power of each individual segment must be set to at least 50%.

It is enlightening to consider that the Sequential Bayes Factor and Group-Sequential procedures are similarly biased in earlier segments. Early stopping acts as a filter that weeds out moderately sized effects: the only effect sizes that are maintained are the ones that are relatively small or in the wrong direction (e.g., when stopping for futility), or the ones that are really large (e.g., when stopping for efficacy). However, due to the cumulative nature of data collection and analysis, the bias inherent in early terminations tends to be compensated for by studies that terminate at later stages (Fan et al., 2004; Goodman, 2007; Schonbrodt et al., 2017). As more observations are collected, the variability in effect size estimates decreases. In meta-analyses of findings stemming from cumulative sequential procedures (such as the SBF and the GS designs), less weight will be given to the relatively biased findings stemming from earlier terminations and – because of their larger sample sizes – more weight will be given to the relatively unbiased findings stemming from later terminations. As a result, the meta-analytic bias in the effect size estimates becomes negligible. Unfortunately, this is not the case for the Independent Segments Procedure. For the Independent Segments Procedure, the first $k_{max} - 1$ segments are highly biased due to their low power, which only allows very small or very large effects to filter through. The last segment has slightly higher power because a weaker α is set (although this

increase in power is less substantial than for cumulative procedures, which receive a marked increase in power due to their accumulating sample size as the study progresses), and thus tends to be slightly less biased. However, because each independent segment has equal sample sizes, they are all given equal weight. As a result, the overall bias in the effect size estimates remains severe.

The median effect size estimates across all possible values of δ are shown in Table 4. As expected, the median effect size estimates stemming from fixed-sample hypothesis tests are equal to the true population effect size. The uncorrected median effect size estimates stemming from the Pocock-like and O'Brien-Fleming-like designs deviate only slightly from the true population effect size. When the population effect size is relatively small ($\delta < 0.2$), the Pocock-like and O'Brien-Fleming-like effect size estimates slightly underestimate the true effect; when the population effect size is relatively large ($\delta > 0.4$), the Pocock-like and O'Brien-Fleming-like effect size estimates slightly overestimate the true population effect size. As can be read in the Table, the group-sequential median-unbiased estimator, used to adjust for bias in the effect size estimates, provides a remarkable improvement in the accuracy of the parameter estimates. In all cases, the group-sequential bias-adjusted effect size estimate is equal to or no more than $d = 0.02$ removed from the true population parameter.

For the unadjusted Bayesian effect size estimates, a similar patterns holds: smaller effect sizes are slightly underestimated, while larger effect sizes are slightly overestimated. The bias-adjusted estimate (the mean of the posterior distribution) increases the estimates for smaller population effect sizes (i.e., for early terminations in which the null was accepted), and shrinks the estimates for larger population effect sizes (i.e., for early terminations in which the alternative was accepted). Although this provides a slight improvement in the parameter estimate for certain

population effect sizes (δ = 0.2, 0.5, and 0.6), the Bayesian bias-corrected estimates do not perform as well as the group-sequential bias-corrected estimates. The mean of the posterior distribution either slightly improves or, for certain parameters, overcorrects the unadjusted Bayesian effect size estimates.

Again, we see that the median effect size estimates stemming from the Independent Segments Procedure deviate most from the true population effect size. Consider the columns of Table 4 that show the median effect size estimates for the respective procedures when the true population effect size δ = 0.2, 0.4, or 0.6. When the true population effect size δ = 0.2, the median effect size estimate on the basis of the ISP is d = 0.05, compared to the following unadjusted (adjusted) estimates for the Pocock-like, SBF, and O'Brien-Fleming-like procedures: 0.13 (0.20), 0.14 (0.17), and 0.16 (0.20). The bias-corrected estimates are no further than d = 0.03 removed from the true population value (with the group-sequential estimates not at all deviating from the true value), while the ISP underestimates the population value by d = 0.15.

When the true population effect size δ = 0.4, the median effect size estimate on the basis of the ISP is d = 0.57, compared to the following unadjusted (adjusted) estimates for the SBF, Pocock-like, and O'Brien-Fleming-like procedures: 0.47 (0.47), 0.45 (0.39), and 0.45 (0.40). The bias-corrected estimates are no further than d = 0.07 removed from the true population value (with the group-sequential estimates deviating only d = 0.01 from the true value), while the ISP overestimates the population value by d = 0.17.

When the true population effect size δ = 0.6, the median effect size estimate on the basis of the ISP is d = 0.74, compared to the following unadjusted (adjusted) estimates for the Pocock-like, SBF, and O'Brien-Fleming-like procedures: 0.67 (0.58), 0.65 (0.62), and 0.64 (0.59). The

bias-corrected estimates are no further than $d = 0.02$ removed from the true population value, while the ISP overestimates the population value by $d = 0.14$.

In our simulations, we find that – for moderate population parameters ($0.4 < \delta < 0.6$) – the ISP (powered to detect an effect of $\delta = 0.5$, with $1 - \beta = 0.8$) severely overestimates effect sizes. In Figure A4 in Appendix A, the reader can find that this observation generalizes across many different specifications of statistical power and hypothesized effect size. Due to the low power of each independent segment, the critical effect size is higher – thus, observed significant effects are likely to exaggerate the true population parameter. In Figure A4, we find that this exaggeration is severe. For example, when $\delta = 0.6$ and power is 80%, the probability of observing a significant effect that is more than 1.5 times the size of the true population parameter (i.e., the probability of observing $d > 0.9$) is 45%, compared to only 15% for a fixed sample hypothesis test.

Figure 7 visualizes the information captured by Table 4, illustrating the differences between the median estimated effect sizes and the true effect sizes. For the Sequential Bayes Factor, we can see that the uncorrected estimated slightly underestimate the true population effect size when it is small, and slightly overestimate the true population effect size when it is large. The bias-adjusted estimate, the mean of the posterior distribution, corrects for this phenomenon – but, as noted before, the Bayesian bias-adjusted estimates only provide a slight improvement or, for certain parameters, an overcorrection. Contrast this with the two bottom row panels for the group-sequential designs. Here, we witness that the group-sequential median unbiased estimator lives up to its name. As before, the ISP shows the greatest overall deviation from the true effect size.

Figure 8 shows the mean squared error of the five procedures for the naive effect size estimates and the available bias-adjusted estimates. The mean squared error is a measure of the accuracy of an estimator, composed of two elements: variance (the spread of effect size estimates from one sample to the other) and squared bias (the difference between the mean estimated effect size and the true population effect size). We find that the Independent Segments Procedure has the greatest degree of mean squared error. In addition, we find that the error stems mostly from the higher variance of effect size estimates. As discussed before, the variance in the effect size estimates stemming from the ISP is caused by the process of conducting three independent mini-experiments, rather than a cumulative study.

In Figures A2 and A3 in Appendix A, the reader can observe that our findings regarding the bias, variance, and mean squared error of the Independent Segments Procedure generalize across many different test specifications. In the language of decision theory, the ISP is inadmissible: its total squared error risk exceeds that of other hypothesis testing procedures (in our worked example of $\delta = 0.5$, MSE = 0.13 for the ISP, compared to 0.08 for the Pocock-like design, 0.06 for the O'Brien-Fleming design, 0.05 for the SBF, and 0.04 for the fixed-sample procedure).

To provide more meat to the argument that the error of the ISP is caused by the process of conducting three independent mini-experiments, we separate the error of the procedures by segment in Figure 9. Here, we can clearly see that stopping after a first look causes error for all sequential procedures. For both the SBF and the group-sequential procedures, however, error becomes near-negligible in the second and third segment. As a result, overall error (across all segments) is much lower. This is akin to the bias compensation process we have described above. For the Independent Segments Procedure, as we now know, error remains substantial in every

segment, leading to non-negligible overall error. We can see that the error for the first $k_{max} - 1$ segments is highest, while the error for the last segment is slightly decreased. As discussed before, this is due to the slightly higher power of the last segment, for which a weaker α is set. This error reduction, however, simply does not compare to the reduction that would have been obtained had the experiment been cumulative.

Somewhat paradoxically, here we have an efficient hypothesis testing procedure that is inefficient in statistical parlance. An efficient estimator is one that has small variance (Fisher, 1922). The ISP, which purports to increase efficiency, does just the opposite: the process of conducting independent mini-experiments and discarding a region of inconclusive data points, while maintaining only the most extreme data points, causes substantial variability in effect size estimates. More importantly, the region of inconclusive data points that the researcher is forced to discard are just those that are closest to the true effect size. As a result, it is near impossible to obtain a credible effect size estimate from the Independent Segments Procedure. We are forced to conclude that the Independent Segments Procedure is a novel method for efficiently rejecting the most plausible effect sizes.

Discussion

In this final section, I present several practical considerations for researchers who wish to choose between the sequential procedures under discussion. Group-sequential designs, Sequential Bayes Factors, and the Independent Segments Procedure each have advantages and disadvantages. In deciding between sequential hypothesis testing procedures, researchers need to carefully weigh the advantages and disadvantages of the procedures of interest, with regard to the context in

which they are conducting their research (Stefan et al., 2020). Table 1 presents several of the commonalities and differences between the sequential hypothesis testing procedures.

A great advantage to the procedures under discussion is that they can all be implemented in a group-wise manner, with a guaranteed upper limit on sample size. Continuous assessment of data can be a serious practical burden. In the field of clinical trial design, early proposals for fully sequential methods that require analysis after every observation – such as the Sequential Probability Ratio Test – did not receive widespread acceptance (Pramanik, Johnson, & Bhattacharya, 2021). It was only when methods were extended to more realistic settings, and allowed for the evaluation of outcomes after groups of participants were observed, that sequential methods became widely used (Jennison & Turnbull, 2000; Pramanik et al., 2021).

Because the O'Brien-Fleming procedure uses very conservative stopping boundaries at earlier looks, the decision rule at the last stage of the experiment (if reached) will be very similar to that of the fixed-sample hypothesis test. In our worked example in Figure 1B, we see that the boundary for rejecting H_0 at the final look is $\alpha_{1,3}$ = .052. To practitioners, this has turned out to be an attractive feature (Jennison & Turnbull, 2000). When using the O'Brien-Fleming correction, it is recommended to collect more than $\frac{1}{k_{max}}$% of the observations before taking the first look, to increase the probability of stopping for efficacy.

Although the ISP has some advantages, none of these advantages are unique. Efficiency and explicit error control can be achieved through alternative methods—none of which pose the disadvantages unique to the ISP. The ISP applies only to settings in which a researcher wishes to test a single null hypothesis and does not desire a principled decision rule to reject an alternative hypothesis. In addition, users of the procedure must guarantee an exact number of participants,

allocated equally across segments. In realistic settings, however, group sizes and increments in information tend to be unequal and unpredictable (Jennison & Turnbull, 2000). Most importantly, the researcher in question should be willing to discard valuable data and abstain from the ability to obtain a credible estimate of effect size.

Miller and Ulrich (2020) present the independent nature of data collection and analysis built-in to the ISP as an attractive feature that creates simplicity and generality. Based on our simulations, however, we find that the information loss inherent to the procedure is a bug, not a feature, that leads to severe difficulty in effect size estimation. Users of the Independent Segments Procedure are advised to choose their α_{strong} and overall statistical power such that the effect size of interest will not be systematically excluded. Ideally, each individual segment should be powered to at least 50% to detect the hypothesized effect size; otherwise, the critical effect size (the effect size needed to reject the null and end the experiment) will be higher than the hypothesized effect size, and the user will thus systematically discard the effect size which they intended to detect. Increasing the power of the individual segments likely means, however, that the user will have to power the overall procedure to well over 90% (see Figure A1 in the Appendix). Consequently, much of the desired efficiency gains will be lost. Given the existence of appealing alternatives that have the same advantages as the ISP (i.e., the ability to efficiently test hypotheses with explicit error control), without its associated disadvantages (i.e., the severe difficulty in effect size estimation), the reader is advised to think twice before utilizing the Independent Segments Procedure.

Earlier, we introduced the reader to two trade-offs of potential importance to the sequential hypothesis tester: long-run error control vs. short-run evidence and efficiency in hypothesis testing vs. accuracy in parameter estimation, respectively. The short-run conception of

evidence on which the (Sequential) Bayes Factor is based, has been accused of throwing out the error control baby with the bathwater (Mayo, 2016: p.2). I show here – in line with previous work – that with appropriate design choices, the SBF procedure (implemented with a lower degree of flexibility than in previous investigations) can have comparable error rates to its frequentist alternatives. When using the Sequential Bayes Factor procedure, however, researchers are recommended to run simulations (e.g., using the R package `BFDA`, Schönbrodt & Wagenmakers, 2018) to choose an optimal design. As discussed earlier, the SBF allows researchers to specify their uncertainty regarding the effect size by means of a prior distribution. Thus, for researchers who wish to incorporate prior beliefs (e.g., because they work in a subject area with high uncertainty about expected effect sizes), the SBF presents an appealing research strategy (A. Stefan et al., 2020). On the other hand, if a researcher wants robust (i.e., guaranteed) error control, group-sequential designs present an appealing research strategy.

Despite the trade-off between accurate effect size estimation and efficient hypothesis testing (Goodman, 2007; Schonbrodt et al., 2017), I find that both the SBF procedure and the GS procedure can provide relatively accurate estimates of effect size. For Sequential Bayes Factors, the mean of the posterior distribution shrinks the effect size estimate in case of early terminations that rejected the null. However, the mean of the posterior distribution does not always provide an improvement: at times, the adjusted estimate is further removed from the true value than the unadjusted estimate. In the case of group-sequential designs, however, we find a remarkable increase in the accuracy of parameter estimation when relying on available median unbiased estimators. Thus, users of group-sequential designs are recommended to always use bias-adjusted estimates for parameter estimation. Overall, the Sequential Bayes Factor and group-sequential procedures can substantially increase research efficiency at only a slight cost to our ability to estimate an effect size.

Conclusion

We have learned that scientific experiments can require substantial time, money, and effort. At times, the well-being of research subjects, human or non-human, is at stake (A. Stefan et al., 2020). In these contexts, sequential hypothesis testing can provide substantial efficiency gains. For researchers who wish to utilize the benefits of sequential hypothesis testing while retaining their ability to obtain a credible effect size estimate, group-sequential designs and Sequential Bayes Factor are an especially good fit. Sequential hypothesis testing procedures are valuable tools for psychological scientists to add to their statistical toolbox.

References

Analytics, R., & Weston, S. (2020). *Iterators: Provides iterator construct*. Retrieved from https://CRAN.R-project.org/package=iterators

APA. (2019). *Publication Manual of the American Psychological Association* (Seventh edition). Washington: American Psychological Association.

Auguie, B. (2017). *gridExtra: Miscellaneous functions for "grid" graphics*. Retrieved from https://CRAN.R-project.org/package=gridExtra

Aust, F. (2019). *Citr: 'RStudio' add-in to insert markdown citations*. Retrieved from https://github.com/crsh/citr

Aust, F., & Barth, M. (2020). *papaja: Prepare reproducible APA journal articles with R Markdown*. Retrieved from https://github.com/crsh/papaja

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The Rules of the Game Called Psychological Science. *Perspectives on Psychological Science*, *7*(6), 543–554. https://doi.org/10.1177/1745691612459060

Barth, M. (2021). *tinylabels: Lightweight variable labels*. Retrieved from https://github.com/mariusbarth/tinylabels

Bates, D., & Maechler, M. (2021). *Matrix: Sparse and dense matrix classes and methods*. Retrieved from https://CRAN.R-project.org/package=Matrix

Beffara Bret, B., Beffara Bret, A., & Nalborczyk, L. (2018). *A fully automated, transparent, reproducible, and blind protocol for sequential analyses* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/v7xpg

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

Borchers, H. W. (2021). *Pracma: Practical numerical math functions*. Retrieved from https://CRAN.R-project.org/package=pracma

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Corporation, M., & Weston, S. (2020). *doParallel: Foreach parallel adaptor for the 'parallel' package*. Retrieved from https://CRAN.R-project.org/package=doParallel

Dienes, Z. (2011). Bayesian Versus Orthodox Statistics: Which Side Are You On? *Perspectives on Psychological Science*, *6*(3), 274–290. https://doi.org/10.1177/1745691611406920

Dodge, H. F., & Romig, H. G. (1929). A Method of Sampling Inspection. *Bell System Technical Journal*, *8*(4), 613–631. https://doi.org/10.1002/j.1538-7305.1929.tb01240.x

Fan, X., DeMets, D. L., & Lan, K. K. G. (2004). Conditional Bias of Point Estimates Following a Group Sequential Test. *Journal of Biopharmaceutical Statistics*, *14*(2), 505–530. https://doi.org/10.1081/BIP-120037195

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *222*(594-604), 309–368. https://doi.org/10.1098/rsta.1922.0009

Fleming, T. R., Harrington, D. P., & O'Brien, P. C. (1984). Designs for group sequential tests. *Controlled Clinical Trials*, *5*(4), 348–361. https://doi.org/10.1016/S0197-2456(84)80014-8

Goodman, S. (1999). Toward Evidence-Based Medical Statistics. 1: The P Value Fallacy. *Annals of Internal Medicine*, *130*(12), 995. https://doi.org/10.7326/0003-4819-130-12-199906150-00008

Goodman, S. (2007). Stopping at Nothing? Some Dilemmas of Data Monitoring in Clinical Trials. *Annals of Internal Medicine*, *146*(12), 882. https://doi.org/10.7326/0003-4819-146-12-200706190-00010

Goodman, S., & Royall, R. (1988). Evidence and scientific research. *American Journal of Public Health*, *78*(12), 1568–1574. https://doi.org/10.2105/AJPH.78.12.1568

Henry, L., & Wickham, H. (2020). *Purrr: Functional programming tools*. Retrieved from https://CRAN.R-project.org/package=purrr

Jennison, C., & Turnbull, B. W. (2000). *Group sequential methods with applications to clinical trials*. Boca Raton: Chapman & Hall/CRC.

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

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

Lakens, D. (2017). Equivalence tests: A practical primer for t-tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *1*, 1–8. https://doi.org/10.1177/1948550617697177

Lakens, D. (2021a). Invited commentary: Comparing the independent segments procedure with group sequential designs. *Psychological Methods*, *26*(4), 498–500. https://doi.org/10.1037/met0000400

Lakens, D. (2021b). The Practical Alternative to the p Value Is the Correctly Used p Value. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *16*(3), 639–648. https://doi.org/10.1177/1745691620958012

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269. https://doi.org/10.1177/2515245918770963

Lan, K. K. G., & Demets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika*, *70*(3), 659–663. https://doi.org/10.1093/biomet/70.3.659

Lawrence Livermore National Laboratory. (2016). *All About that Bayes: Probability, Statistics, and the Quest to Quantify Uncertainty*.

Mayo, D. (2016). Don't Throw Out the Error Control Baby With the Bad Statistics Bathwater: A Commentary. *The American Statistician*, *70*(Online Discussion), 1–2.

Mayo, D. (2018). *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781107286184

McGrath, J. E. (1981). Dilemmatics: The Study of Research Choices and Dilemmas. *American Behavioral Scientist*, *25*(2), 179–210. https://doi.org/10.1177/000276428102500205

Microsoft, & Weston, S. (2020). *Foreach: Provides foreach looping construct*. Retrieved from https://CRAN.R-project.org/package=foreach

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

Morey, R. D., & Lakens, D. (2016). *Why most of psychology is statistically unfalsifiable (Version 1.0)*. https://doi.org/10.5281/zenodo.838685

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

Morey, R. D., & Rouder, J. N. (2018). *BayesFactor: Computation of bayes factors for common designs*. Retrieved from https://CRAN.R-project.org/package=BayesFactor

Müller, K., & Wickham, H. (2021). *Tibble: Simple data frames*. Retrieved from https://CRAN.R-project.org/package=tibble

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's Renaissance. *Annual Review of Psychology*, *69*(1), 511–534. https://doi.org/10.1146/annurev-psych-122216-011836

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. Series A, Containing Papers of a Mathematical or Physical Character*, *231*, 289–337.

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News*, *6*(1), 7–11. Retrieved from https://journal.r-project.org/archive/

Pocock, S. J. (1977). Group Sequential Methods in the Design and Analysis of Clinical Trials. *Biometrika*, *64*(2), 191–199. https://doi.org/10.2307/2335684

Pocock, S. J., & Hughes, M. D. (1989). Practical problems in interim analyses, with particular regard to estimation. *Controlled Clinical Trials*, *10*(4), 209–221. https://doi.org/10.1016/0197-2456(89)90059-7

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

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. (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

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*(3), 520–547. https://doi.org/10.1111/tops.12214

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

Schonbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychon Bull Rev*, 15.

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

Schönbrodt, F. D., & Wagenmakers, E.-J. (2018). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*, *25*(1), 128–142. https://doi.org/10.3758/s13423-017-1230-y

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Stefan, A. M., Gronau, Q. F., Schonbrodt, F. D., & Wagenmakers, E.-J. (2019). A tutorial on Bayes Factor Design Analysis using an informed prior. *Behavior Research Methods*, *51*(3), 1042–1058. https://doi.org/10.3758/s13428-018-01189-8

Stefan, A., Schonbrodt, F. D., Evans, N. J., & Wagenmakers, E.-J. (2020). *Efficiency in Sequential Testing: Comparing the Sequential Probability Ratio Test and the Sequential Bayes Factor Test* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/ry4fw

Van Doorn, J., van den Bergh, D., Böhm, U., Dablander, F., Derks, K., Draws, T., … Wagenmakers, E.-J. (2020). The JASP guidelines for conducting and reporting a Bayesian analysis. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-020-01798-5

Wagenmakers, E.-J., & Gronau, Q. F. (n.d.). *A Compendium of Clean Graphs in R*. https://www.shinyapps.org/apps/RGraphCompendium/index.php#prior-and-posterior.

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*(1), 35–57. https://doi.org/10.3758/s13423-017-1343-3

Wald, A. (1973). *Sequential analysis*. New York: Dover Publications.

Wassmer, G., & Brannath, W. (2016). *Group Sequential and Confirmatory Adaptive Designs in Clinical Trials*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-32562-0

Wassmer, G., & Pahlke, F. (2021). *Rpact: Confirmatory adaptive clinical trial design and analysis*. Retrieved from https://CRAN.R-project.org/package=rpact

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

Wickham, H. (2019). *Stringr: Simple, consistent wrappers for common string operations*. Retrieved from https://CRAN.R-project.org/package=stringr

Wickham, H. (2021a). *Forcats: Tools for working with categorical variables (factors)*. Retrieved from https://CRAN.R-project.org/package=forcats

Wickham, H. (2021b). *Tidyr: Tidy messy data*. Retrieved from https://CRAN.R-project.org/package=tidyr

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wickham, H., François, R., Henry, L., & Müller, K. (2021). *Dplyr: A grammar of data manipulation*. Retrieved from https://CRAN.R-project.org/package=dplyr

Wickham, H., & Hester, J. (2021). *Readr: Read rectangular text data*. Retrieved from https://CRAN.R-project.org/package=readr

Table 1

Commonalities and Differences between Three Sequential Hypothesis Testing Procedures

	Group-Sequential Designs	Independent Segments Procedure	Sequential Bayes Factor Design
Aim	Efficiently reject H_0 in favor of an alternative H_1	Efficiently reject H_0 in favor of an alternative H_1	Efficiently examine the evidence for H_0 relative to an alternative H_1, or vice versa
Nature of data	Cumulative	Independent	Cumulative
Flexibility	Moderately flexible	Highly restrictive	Highly flexible
Principled rejection of the alternative	Possible	Impossible	Possible
Control of error rates	Direct	Direct	Indirect (e.g., through simulation)
Bias-adjusted estimator	Available	Does not exist	Available
Consideration of prior beliefs	Absent	Absent	Present and required

Table 2A

True Negative, False Positive, and Inconclusive Rate

Procedure	True Negative	Inconclusive	False Positive
Fixed	0.79	0.16	0.05
Pocock	0.88	0.07	0.05
O'Brien-Fleming	0.87	0.08	0.05
Bayes	0.85	0.09	0.06
ISP	0.90	0.05	0.05

Note. Error rates of the procedures when the true population effect size $\delta = 0$.

Table 2B

True Positive, False Negative, and Inconclusive Rate

Procedure	True Positive	Inconclusive	False Negative
Fixed	0.80	0.09	0.10
Pocock	0.80	0.05	0.15
O'Brien-Fleming	0.82	0.06	0.13
Bayes	0.84	0.08	0.08
ISP	0.80	0.03	0.17

Note. Error rates of the procedures when the true population effect size $\delta = 0.5$.

Table 3

Average Sample Size Required

Procedure	$d = -0.2$	$d = 0$	$d = 0.2$	$d = 0.4$	$d = 0.5$	$d = 0.6$	$d = 0.8$	$d = 1$
Fixed	51	51	51	51	51	51	51	51
Pocock	26	32	39	41	39	35	29	24
O'Brien-Fleming	30	34	41	43	41	39	33	30
Bayes	30	39	47	46	42	37	30	26
ISP	28	33	41	45	42	39	31	27

Note. Average sample size required for five hypothesis testing procedures across a range of unexpected effect sizes. All frequentist procedures were powered to detect an effect of $\delta = 0.5$. For the Sequential Bayes Factor procedure, simulations were run to choose an optimal sample size, decision threshold, and prior distribution, with the goal of approaching similar error rates as the frequentist procedures.

Table 4

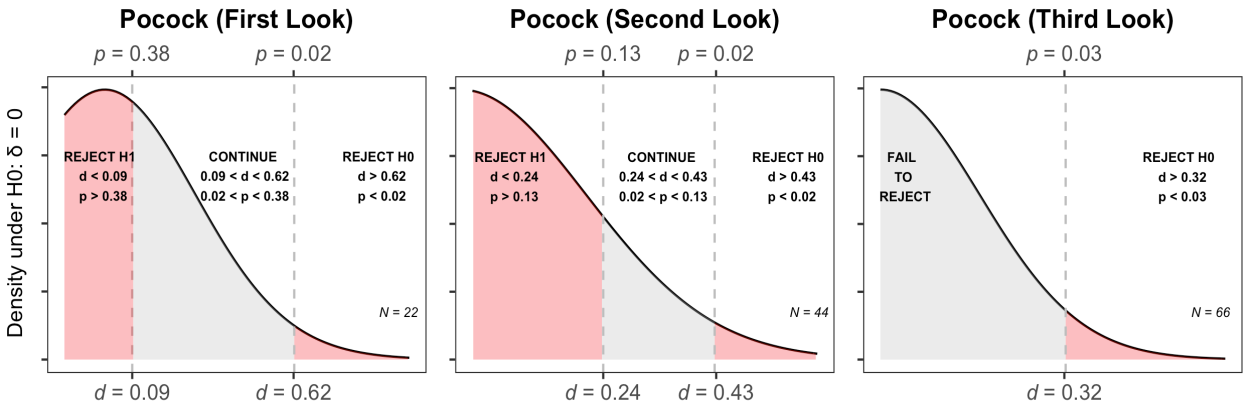
Median Effect Size Estimate

Procedure	$d = -0.2$	$d = 0$	$d = 0.2$	$d = 0.4$	$d = 0.5$	$d = 0.6$	$d = 0.8$	$d = 1$
Fixed	-0.20	0.00	0.20	0.40	0.50	0.60	0.80	1.00
Pocock	-0.21	-0.04	0.13	0.45	0.56	0.67	0.83	1.01
Pocock adjusted	-0.20	0.00	0.20	0.39	0.49	0.58	0.80	1.00
O'Brien-Fleming	-0.20	-0.02	0.16	0.45	0.53	0.64	0.83	1.01
O'Brien-Fleming adjusted	-0.20	0.00	0.20	0.40	0.49	0.59	0.80	1.00
Bayes	-0.21	-0.04	0.14	0.47	0.57	0.65	0.83	1.03
Bayes adjusted	-0.07	0.05	0.17	0.47	0.55	0.62	0.75	0.89
ISP	-0.23	-0.08	0.05	0.57	0.66	0.74	0.87	1.03

Note. The table shows the median effect size estimate across simulations. The naive effect size measure is Cohen's *d*. For the group-sequential designs, a median unbiased estimator (available in statistical software such as `rpact`, Wassmer & Pahlke, 2020) was used to calculate the bias-adjusted estimate; for the Sequential Bayes Factor, the bias-adjusted estimate is the mean of the posterior distribution.

Figure 1A

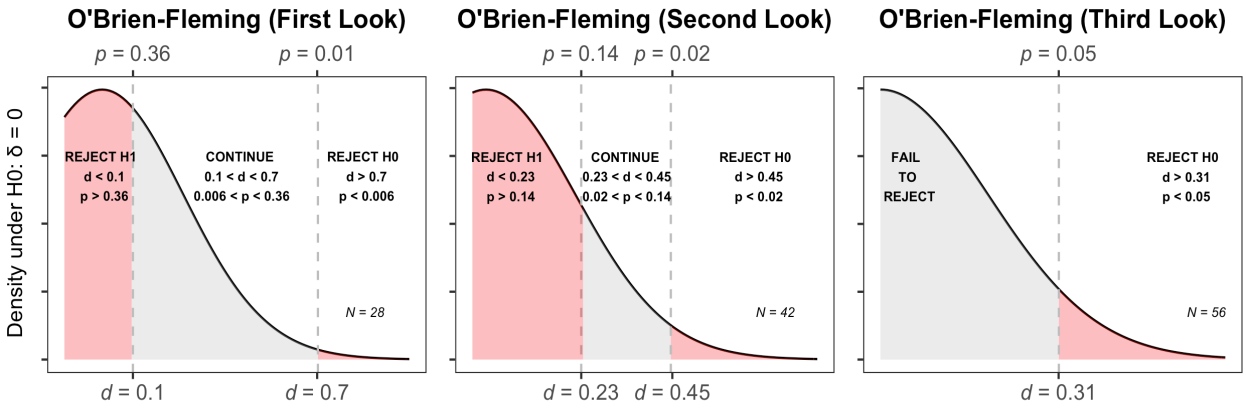
Pocock-like Group-Sequential Design: A worked example



Note. Pocock-like group-sequential design for a one-tailed, two-sample *t* test with α = .05, β = .20 and k_{max} = 3 equally spaced looks. Critical values are determined on the basis of Pocock-like α- and β-spending functions with binding futility bounds.

Figure 1B

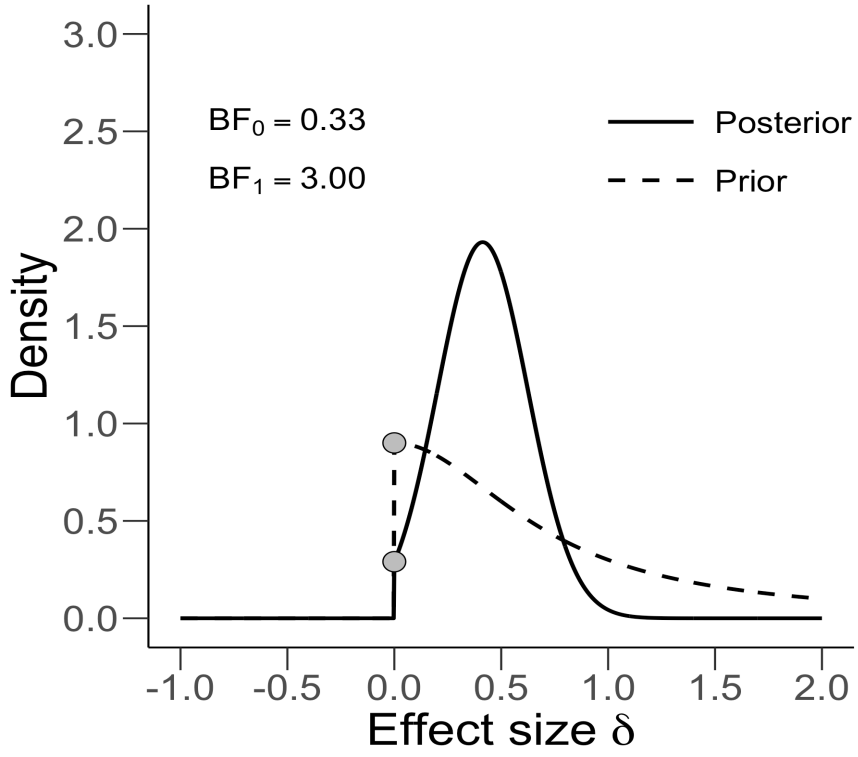
O'Brien-Fleming-like Group-Sequential Design: A worked example



Note. O'Brien-Fleming-like group-sequential design for a one-tailed, two-sample *t* test with α = .05, β = .20 and k_{max} = 3 unequally spaced looks (the looks were set to 50%, 75%, and 100% of the data, respectively). Critical values are determined on the basis of O'Brien-Fleming-like α- and β-spending functions with binding futility bounds.

Figure 2A

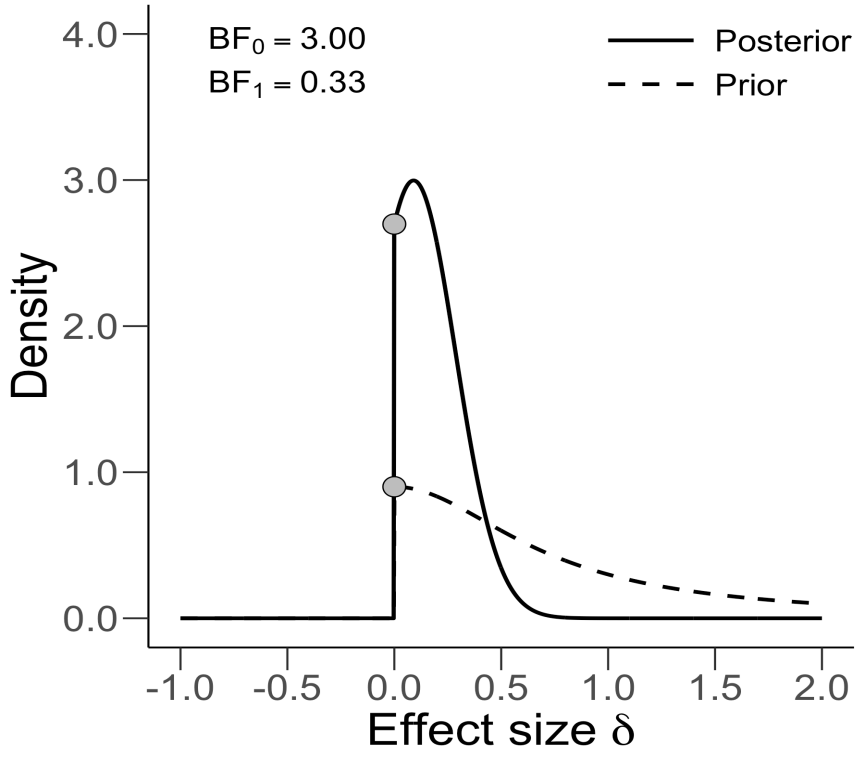
Bayes Factor in Favor of the Alternative Hypothesis



Note. BF_1 = 3 indicates that the alternative hypothesis (H_1: $\delta > 0$) performs three times better at predicting the observed data than the null hypothesis (H_0: $\delta = 0$). Figure based on Jeffreys' Amazing Statistics Program (JASP), adapted from Wagenmakers & Gronau (n.d.)

Figure 2B

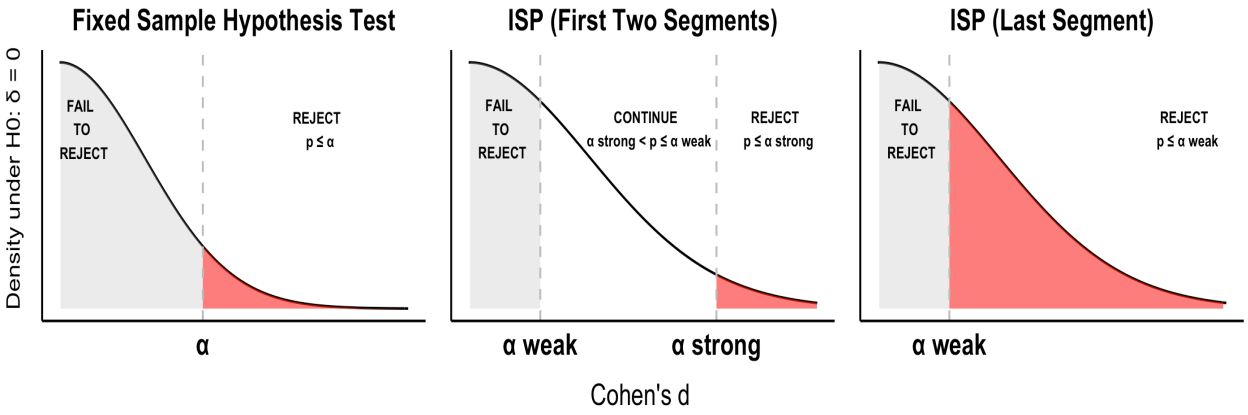
Bayes Factor in Favor of the Null Hypothesis



Note. BF_0 = 3 indicates that the null hypothesis (H_0: $\delta = 0$) performs three times better at predicting the observed data than the alternative hypothesis (H_1: $\delta > 0$). Figure based on JASP, adapted from Wagenmakers & Gronau (n.d.)

Figure 3A

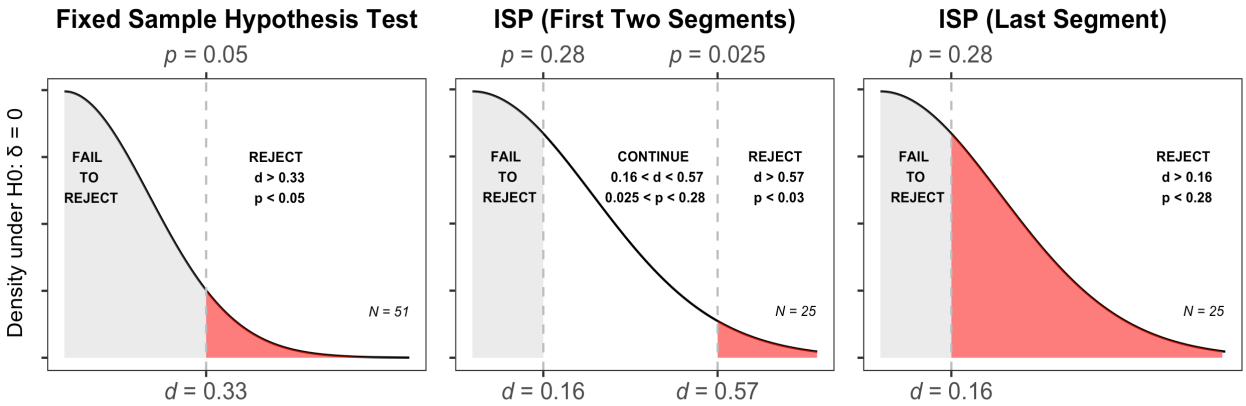
Fixed Sample versus Independent Segments Hypothesis Testing: A conceptual overview



Note. A general comparison of the fixed sample hypothesis test and the independent segments procedure.

Figure 3B

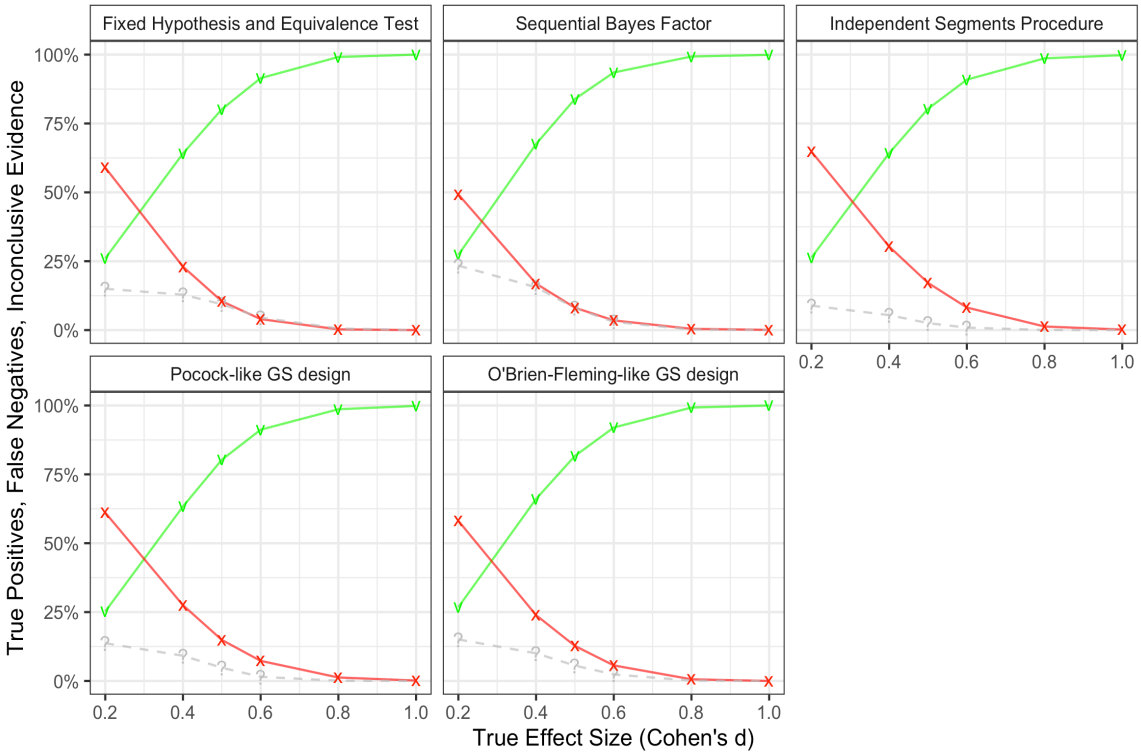
Fixed Sample versus Independent Segments Hypothesis Testing: A worked example



Note. Specific example of fixed sample versus independent segments one-tailed, two-sample *t* tests using $\alpha = .05$ and $1 - \beta = 0.8$ for $\delta = 0.5$.

Figure 4

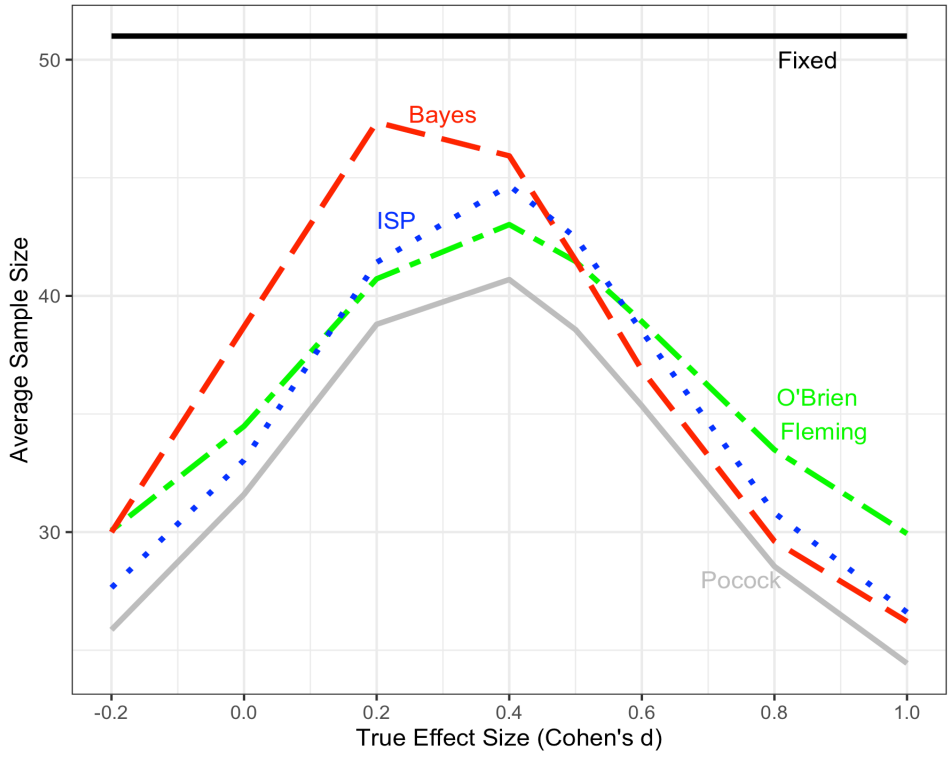
Error rates of the four procedures



Note. Red lines with crosses indicate the rates of false negative results; green lines with check marks indicate the rates of true positive results; grey, dashed lines with question marks indicate the rates of inconclusive results.

Figure 5

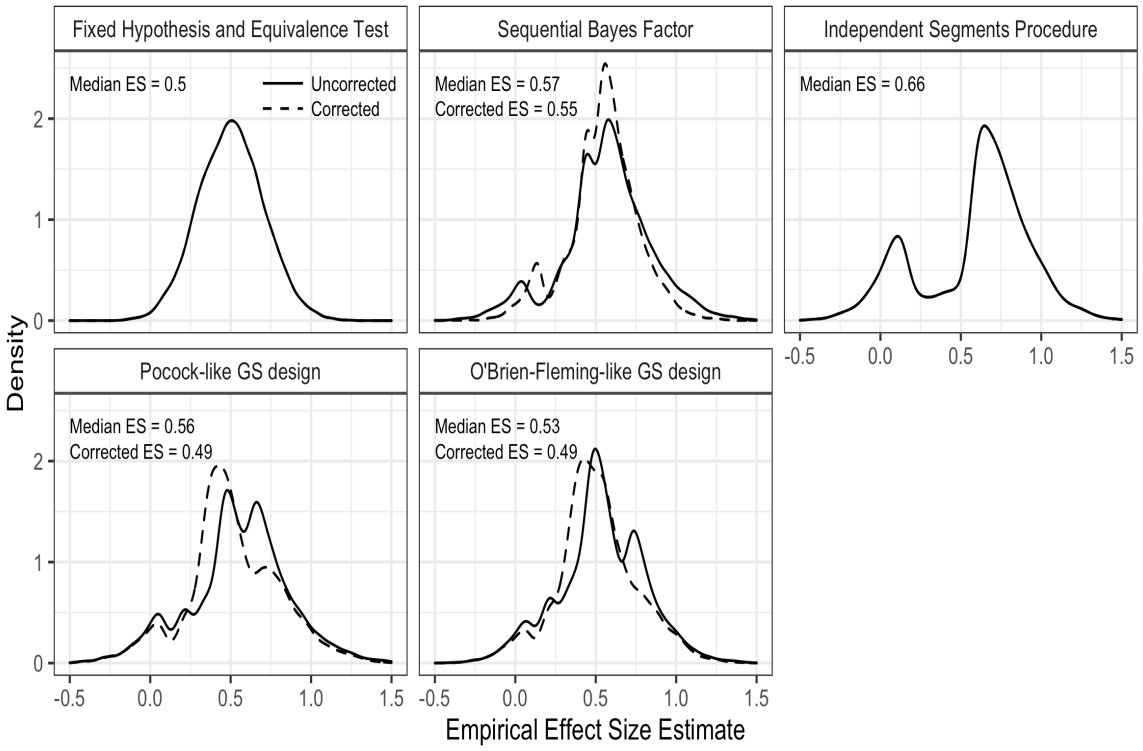
Efficiency of the four procedures



Note. Average sample size required across a range of true effect sizes. The solid, black line represents the fixed-sample Neyman-Pearson hypothesis test; the long-dashed, red line represents the Sequential Bayes Factor procedure; the two-dashed, green line represents the O-Brien-Fleming-like GS procedure; the solid, grey line represents the Pocock-like GS procedure; and the dotted, blue line represents the Independent Segments Procedure.

Figure 6

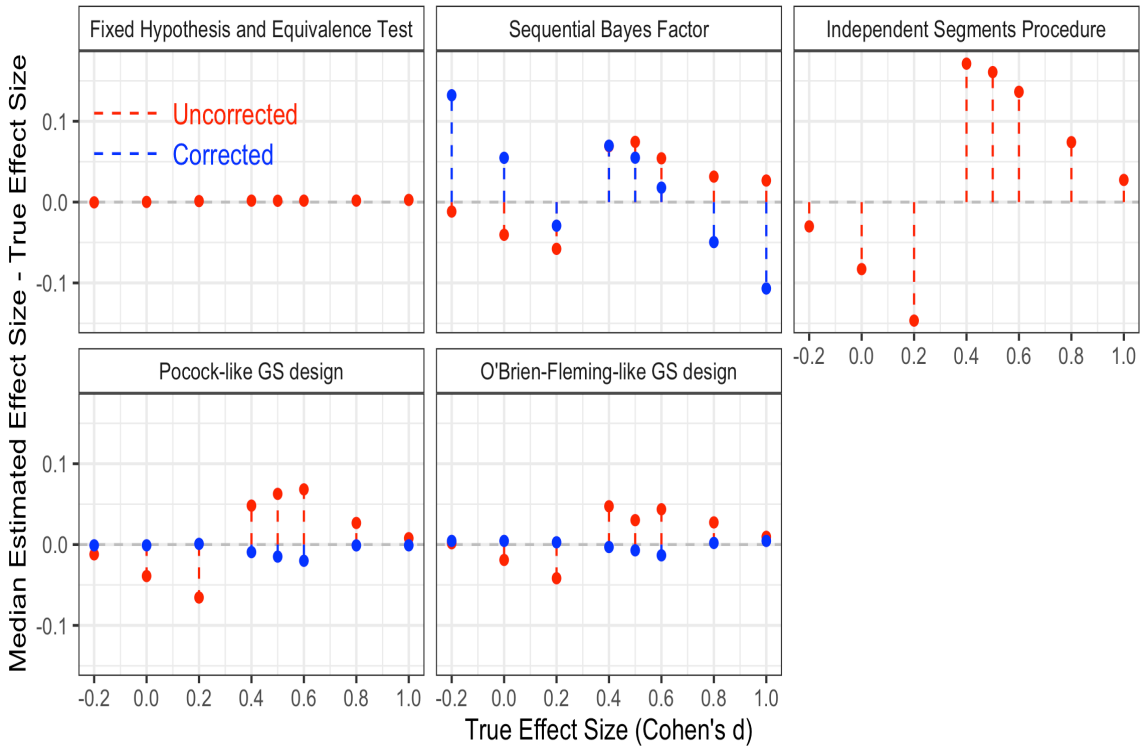
Distribution of Empirical Effect Size Estimates



Note. One-tailed, two-sample *t* tests using, for the frequentist procedures, $\alpha = .05$ and $1 - \beta = 0.8$ for $\delta = 0.5$. *N*s per group per look are 51, 25, and 22 for the Fixed Sample, Independent Segments, and Pocock-like hypothesis tests. For the O'Brien Fleming test, $N = 28$ per group for the first look, and 14 per group for the last two looks. All sequential procedures were set to a maximum number of $k_{max} = 3$ looks. For the Sequential Bayes Factor, data was collected in (at most three) batches of 22 subjects per group. Figure includes the median effect size estimate (the solid line is the naive estimate; the dashed line is the bias-adjusted estimate). The true effect size $\delta = 0.5$.

Figure 7

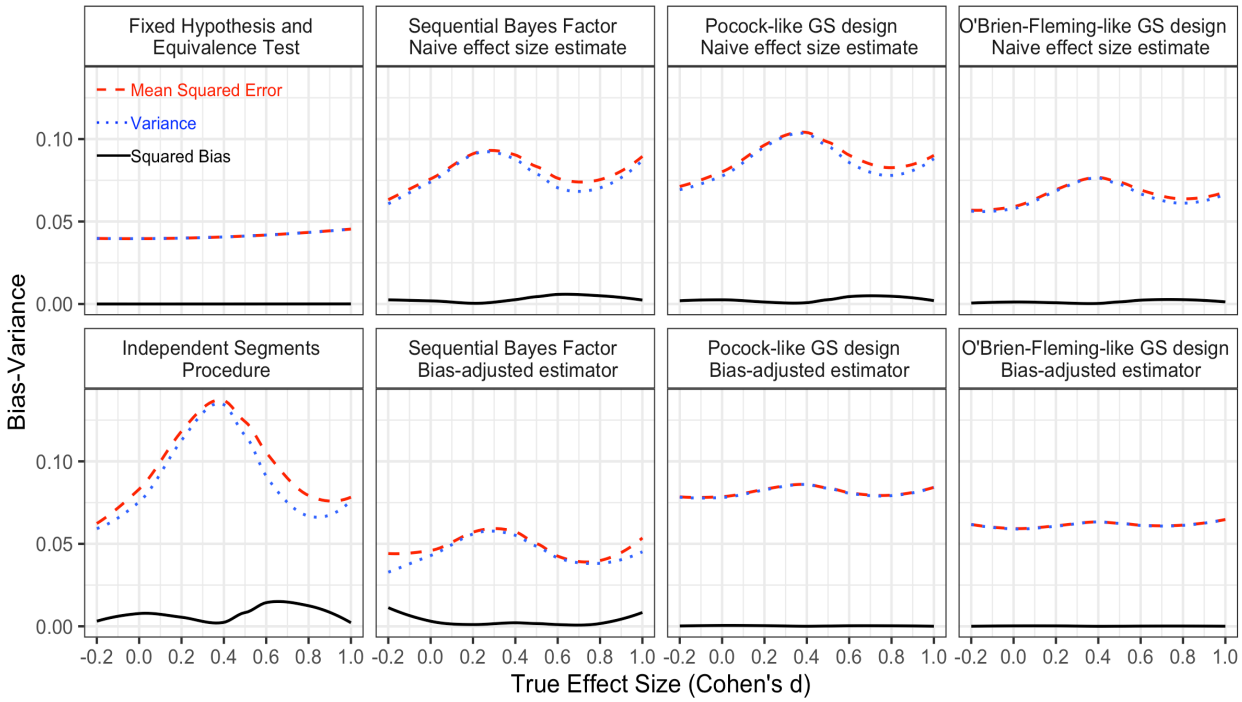
Median Bias in Empirical Effect Size Estimates



Note. Figure displays the median estimated effect size minus the true effect size for the five different procedures. The median uncorrected effect size estimate is shown in red; the median bias-adjusted effect size estimate, if available, is shown in blue.

Figure 8

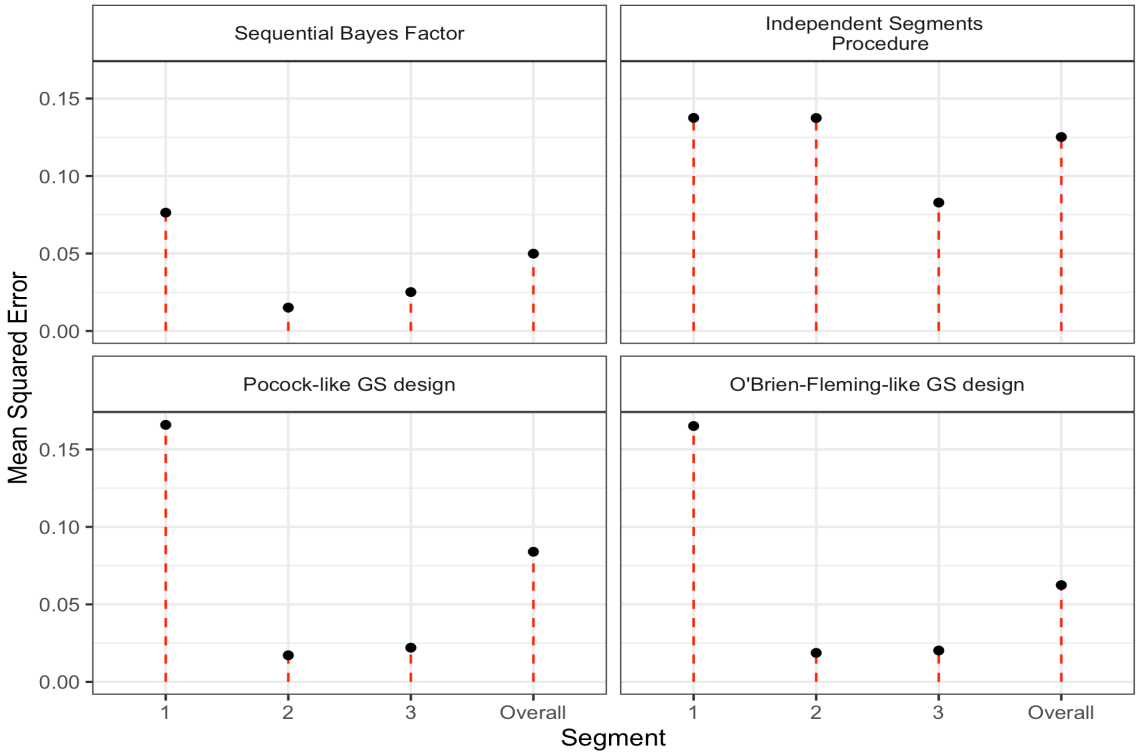
Bias-Variance in Empirical Effect Size Estimates



Note. Figure displays the squared bias, variance, and mean squared error (the sum of the squared bias and the variance) in empirical effect size estimates stemming from the five procedures. Results are split into two panels for the three procedures that have bias corrections available. For these three procedures – the Sequential Bayes Factor, Pocock-like group-sequential design, and O'Brien-Fleming-like group-sequential design – the top panel shows the bias and variance in the naive effect size estimates; the bottom panel shows the bias and variance in the bias-adjusted estimates.

Figure 9

Mean Squared Error in Empirical Effect Size Estimates by Segment

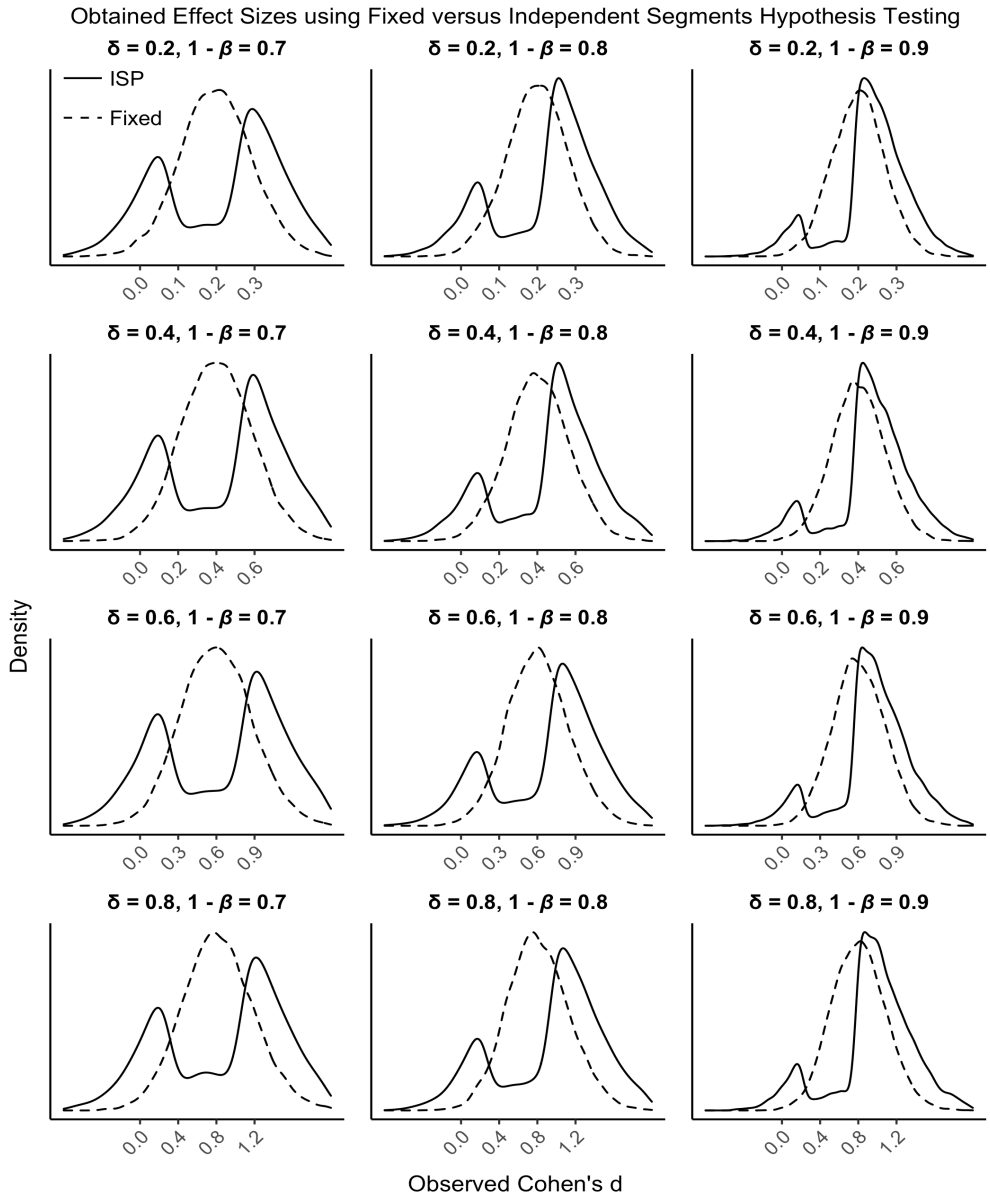


Note. Figure displays the mean squared error in empirical effect size estimates by segment, and the overall mean squared error across all three segments. For the SBF and GS designs, corrected effect size estimates were used to calculate the MSE.

Appendix A

Figure A1

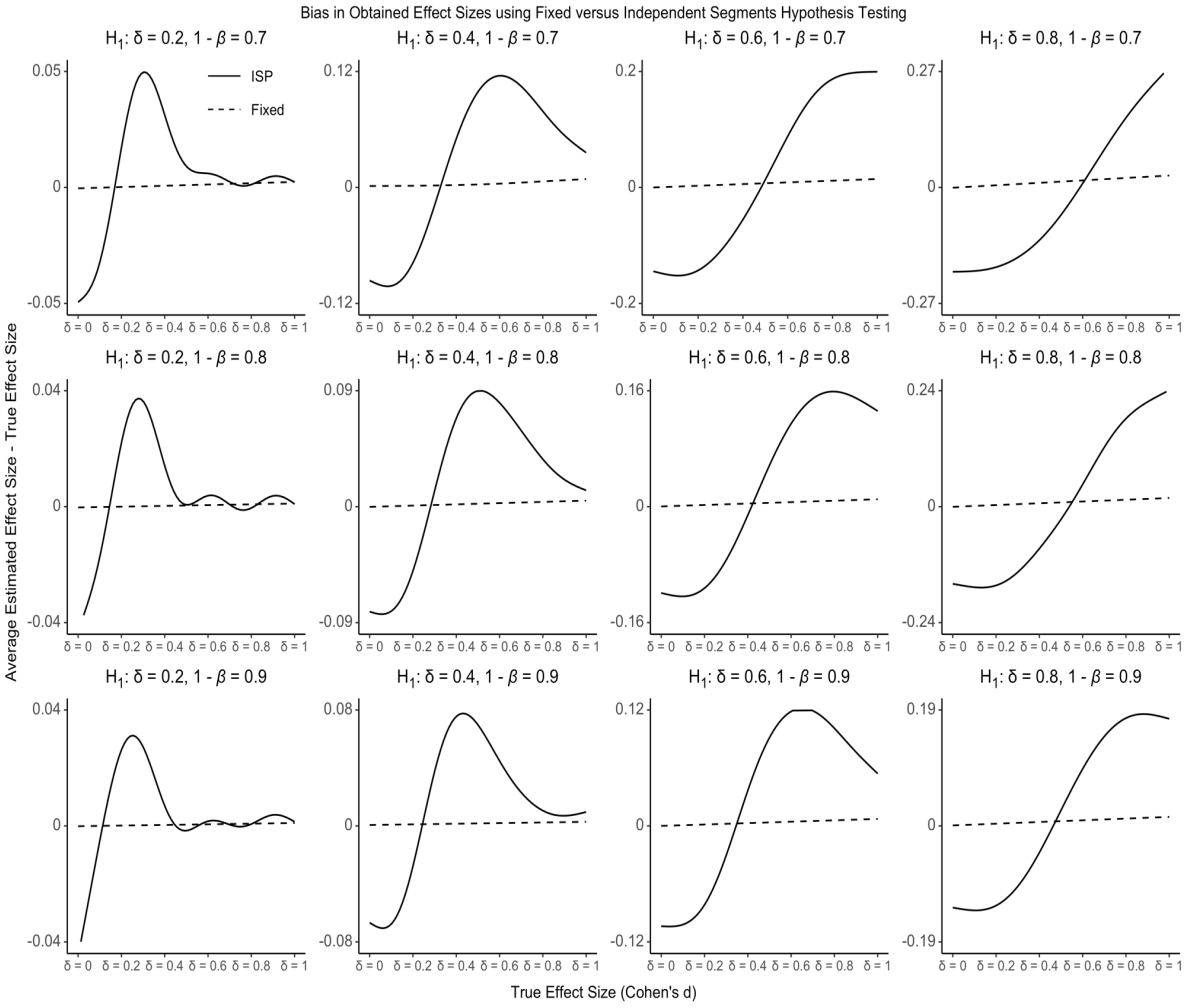
Distribution of Effect Sizes Using Fixed versus Independent Segments Hypothesis Testing



Note. Distributions of effect sizes for the ISP (compared to fixed-sample hypothesis testing), across twelve different combinations of statistical power ($1 - \beta$ = 0.7, 0.8, and 0.9) and population effect size (δ = 0.2, 0.4, 0.6, 0.8).

Figure A2

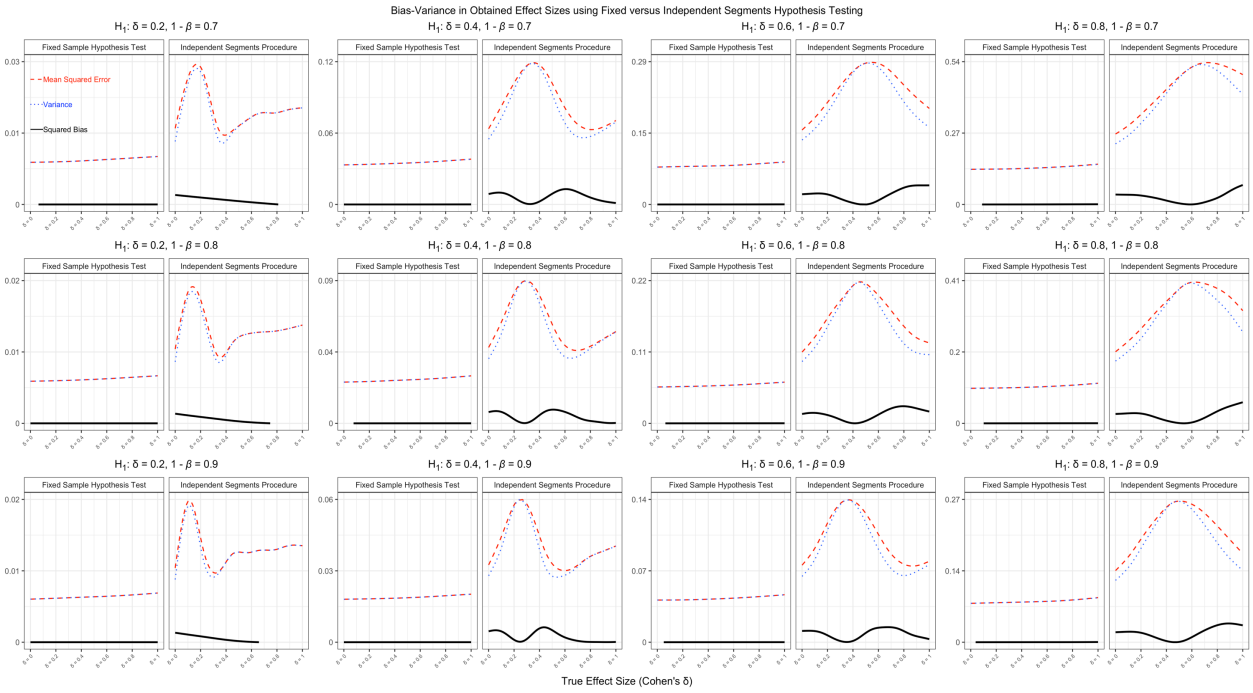
Bias in Obtained Effect Sizes Using Fixed versus Independent Segments Hypothesis Testing



Note. Bias in obtained effect sizes for the ISP (compared to fixed-sample hypothesis testing), across twelve different combinations of statistical power ($1 - \beta$ = 0.7, 0.8, and 0.9) and hypothesized population effect size ($H_1: \delta$ = 0.2, 0.4, 0.6, 0.8), for a range of true population effect sizes ($0 \leq \delta \leq 1$).

Figure A3

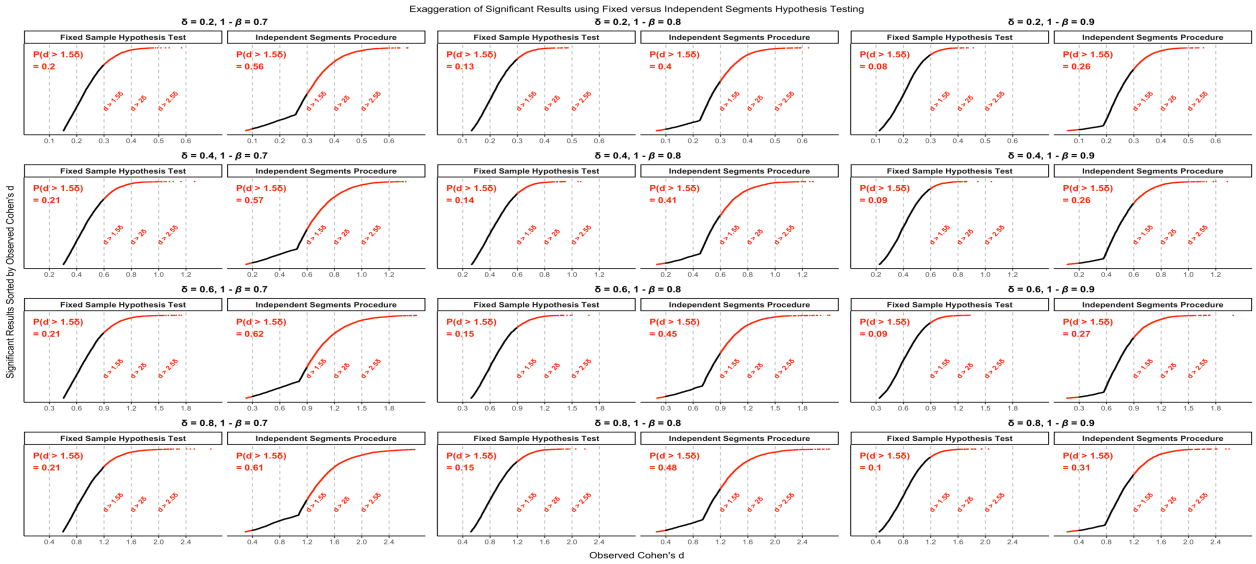
Bias-Variance in Effect Sizes Using Fixed versus Independent Segments Hypothesis Testing



Note. Bias and variance in obtained effect sizes for the ISP (compared to fixed-sample hypothesis testing), across twelve different combinations of statistical power ($1 - \beta = 0.7$, 0.8, and 0.9) and hypothesized population effect size ($H_1: \delta = 0.2$, 0.4, 0.6, 0.8), for a range of true population effect sizes ($0 \leq \delta \leq 1$).

Figure A4

Exaggeration of Significant Results using Fixed versus Independent Segments Hypothesis Testing



Note. The figure shows obtained significant results for the two different hypothesis testing procedures, across twelve different combinations of statistical power ($1 - \beta = 0.7$, 0.8, and 0.9) and hypothesized population effect size ($H_1: \delta = 0.2$, 0.4, 0.6, 0.8). The number of significant results is held constant across conditions ($N = 5{,}000$), and the significant results are sorted by the observed Cohen's *d*. Results that deviate more than 50% from the true population value are shown in red. Each panel denotes the probability of obtaining a Cohen's *d* that is 50% larger than the true population effect size δ.