A Simple, General, and Efficient Method for Sequential Hypothesis Testing: The Independent Segments Procedure

Jeff Miller

University of Otago

Rolf Ulrich

University of Tübingen

Please address editorial correspondence to: Jeff Miller, Department of Psychology, University of Otago, PO Box 56, Dunedin 9054, New Zealand.

International FAX: 64-3-479-8335.

email: miller@psy.otago.ac.nz.

Version of July 14, 2020, as accepted for publication in *Psychological Methods*.

Not the final version: still to be copy-edited.

Author Note

Address correspondence to Jeff Miller, Department of Psychology, University of Otago, Dunedin, New Zealand, or Rolf Ulrich, Department of Psychology, University of Tübingen, Schleichstr. 4, 72076 Tübingen, Germany. Electronic mail may be sent to miller@psy.otago.ac.nz or ulrich@uni-tuebingen.de.

Abstract

We propose a new sequential hypothesis testing procedure in which data are collected and analyzed in a series of *independent* segments. As in fixed-sample hypothesis testing and in previous sequential procedures, the overall α level can be set to any desired value. Like other sequential procedures, the independent segments procedure generally requires smaller samples than fixed-sample procedures—often approximately 30% smaller—to achieve the same α level and statistical power. Relative to other sequential procedures, the new method has the advantages that it is simpler to use, requires fewer assumptions, and can be used with a wider array of statistical tests. Thus, in some circumstances the independent segments procedure may provide an attractive option for increasing the efficiency of statistical testing.

Keywords: hypothesis testing; sequential sampling; research efficiency;

A Simple, General, and Efficient Method for Sequential Hypothesis Testing: The Independent Segments Procedure

On a global scale, scientific research is a huge enterprise that consumes vast resources (e.g., an estimated US$240 billion for biomedical research in 2009; Røttingen et al., 2013). Partly because of recent reports that many results are not replicable (e.g., Baker, 2016; Camerer et al., 2016; Gorroochurn, Hodge, Heiman, Durner, & Greenberg, 2007; Ioannidis, 2005; Open Science Collaboration, 2015), there is increasing concern that many of these resources are being wasted (e.g., Chalmers et al., 2014; Chalmers & Glasziou, 2009; Freeman, 2017; Ioannidis et al., 2014; Macleod et al., 2014). The question of how research methodology can be improved is therefore critical, particularly within social, psychological, and biomedical sciences (e.g., Begley & Ioannidis, 2015; Munafò et al., 2017; Nosek, Spies, & Motyl, 2012).

It has long been known that research efficiency can be improved by using sequential procedures. These differ from the more common fixed-sample designs in that the total sample size for a study is not fixed in advance. Instead, observations are collected and analyzed in successive steps, and data collection is stopped when the evidence satisfies certain criteria (e.g., Pocock, 1977; Rushton, 1950; Wald, 1947; for reviews, see Govindarajulu, 2004 and Schnuerch & Erdfelder, 2020). When they can be used, these designs are known to be more efficient than fixed-sample designs (i.e., they achieve the same statistical power with smaller average sample sizes; Jennison & Turnbull, 2000; Proschan, Lan, & Wittes, 2006; Wetherill, 1975). Unfortunately, these procedures are under-utilized even in the medical research areas where they are most common (Stevely et al., 2015), perhaps because they are difficult for non-specialists to use (Albers, 2019), and they have not been widely adopted in other research areas either, for a variety of reasons (Lakens, 2014).

This article describes a new *independent segments* sequential hypothesis testing procedure with several features that may make it more appealing than previous ones. We will describe this approach in the following section and subsequently contrast it with previous sequential approaches.

Independent Segments Procedure

The new independent segments approach to hypothesis testing is illustrated in Figure 1. Specifically, a researcher using this procedure tests a single H_0 by conducting a series of at most k_{max} identical and independent data-collection segments as part of a single overall study. After each segment k, an observed p value, $\mathbf{p}_{obs,k}$, is computed from the data of that segment and is compared against two critical values, α_{strong} and α_{weak}, ($\alpha_{strong} < \alpha < \alpha_{weak}$), that are chosen to give the desired *overall* Type I error rate (i.e., α) for the full study as a 1-tailed test of H_0 (see Appendix A for details). After each of the first $k_{max} - 1$ segments, the researcher computes $\mathbf{p}_{obs,k}$ and rejects H_0 if the observed effect is very strong in the tested direction (i.e., if $\mathbf{p}_{obs,k} \leq \alpha_{strong}$) or fails to reject (FTR) H_0 if the effect is very weak or reversed (i.e., if $\mathbf{p}_{obs,k} > \alpha_{weak}$). If either of these two conditions is met, the researcher stops without testing any further segments. If neither condition is met (i.e., $\alpha_{strong} < \mathbf{p}_{obs,k} \leq \alpha_{weak}$), the researcher continues on to the next segment. If the final segment is reached, H_0 is rejected if there is at least a small effect in the hypothesized direction (i.e., if $\mathbf{p}_{obs,k} \leq \alpha_{weak}$).

Figure 2 shows two numerical examples in the form of tree diagrams with three segments for a researcher carrying out a 2-sample t-test using $k_{max} = 3$ independent segments with $n_s = 50$ per segment (i.e., $n = 25$ per group), $\alpha = 0.05$, and $\alpha_{strong} = 0.025$. To achieve an overall Type I error rate of $\alpha = 0.05$, the researcher computes $\alpha_{weak} = 0.28178$, as is described in Appendix A. Figure 2A depicts the possible outcomes and their associated probabilities when H_0 is true. Note that the probabilities of rejecting H_0 after the first, second, and third segments are 0.0250, 0.0064, and 0.0186, respectively. These three probabilities sum to $\alpha = 0.05$ and thus the procedure yields the desired (preselected) 0.05 Type I error rate under H_0. The three lower branches depict the outcomes of failing to reject H_0 with their associated probabilities. These probabilities are 0.7182, 0.1844, and 0.0474 for the first, second, and third segments, and they sum to 0.95, which is the probability of a correct failure to reject H_0. These probabilities can also be used to compute the probability that the procedure would terminate at each segment. This probability is

$0.0250 + 0.7182 = 0.7432$ for the first segment, $0.0064 + 0.1844 = 0.1908$ for the second segment, and $0.0186 + 0.0474 = 0.0659$ for the last segment. Accordingly, when H_0 is true, there is a high probability that the study would terminate after the first or second segment and thereby preserve the resources that would otherwise be consumed by additional sampling, allowing these resources to be devoted to future studies. More specifically, the expected total sample size for this case is
$\mathrm{E}[\mathbf{N}|d = 0] = 50 \cdot 0.7432 + 100 \cdot 0.1908 + 150 \cdot 0.0659 = 66.1$.

Figure 2B depicts a situation where there is a true effect in the hypothesized direction (i.e., H_0 is false). Specifically, we consider the case in which the true group means differ by $d = 0.5$ units relative to the common true standard deviation. Power calculations indicate that the probabilities for rejecting H_0 at each step are now 0.4100, 0.1937, and 0.1969 (see Appendix A for details). These sum to 0.8006, which is the overall statistical power $1 - \beta$ for this case. The termination probabilities are $0.4100 + 0.1176 = 0.5276$ at the first segment, $0.1937 + 0.0556 = 0.2492$ at the second segment, and $0.1969 + 0.0262 = 0.2232$ at the last segment. With $d = 0.5$, then, the expected total sample size is $\mathrm{E}[\mathbf{N}] = 50 \cdot 0.5276 + 100 \cdot 0.2492 + 150 \cdot 0.2232 = 84.8$.

Like all sequential procedures, the independent segments strategy is attractive because it allows researchers to stop when the early results favor a given decision, thereby preserving resources for future studies. One possibility is that the researcher may stop because the results of an early segment do not provide at least weak evidence against H_0 (i.e., $\mathbf{p}_{obs,k} > \alpha_{weak}$). Alternatively, the researcher may stop if the results of an early segment provide strong evidence against H_0 (i.e., $\mathbf{p}_{obs,k} < \alpha_{strong}$). Critically, sequential procedures hold the overall Type I error rate constant at the desired α level while providing such possibilities for early stopping.

By way of comparison, consider a fixed-sample researcher using 2-sample t-tests and $\alpha = 0.05$, like the independent segments researcher shown in Figure 2. To obtain the same statistical power $1 - \beta = 0.8$ for an effect size of $d = 0.5$ as the independent segments researcher, the fixed-sample researcher would need a sample size of $\mathbf{N} = 100$, which is larger than the independent segments researcher's expected sample sizes under

both H_0 and H_1, 66.1 and 84.8, as computed earlier. Thus, the independent segments strategy is more efficient than the fixed-sample one, since it achieves the same level of power with smaller on-average sample sizes.

Comparison With Previous Sequential Approaches

To our knowledge, the independent segments procedure is unique among sequential procedures in analyzing successive data segments independently rather than cumulatively. The advantage of this approach is that it greatly simplifies the computation of key quantities such as α level, power, and expected sample size. Furthermore, the new procedure is unusual in considering only a succession of $\mathbf{p}_{obs,k}$ values, which provides its generality and simplicity. Frick (1998) suggested a superficially similar procedure that also considers only successive $\mathbf{p}_{obs,k}$ values. These values were computed cumulatively rather than independently, however, which rendered his procedure mathematically intractable. Simulations were required to determine key quantities for any desired situation (i.e., statistical test, α level, effect size, etc), which may explain why his procedure has not been widely discussed.

Other sequential procedures have been widely studied, however, and it seems worthwhile to contrast the independent segments approach with these. Although it is beyond the scope of this article to present more than an overview of the many previously-suggested sequential procedures, Table 1 summarizes the major differences between our new approach and two major classes of previous sequential procedures.

One major class of sequential procedures is based on the sequential probability ratio test (SPRT; Wald, 1947), which can be used for deciding between two simple point hypotheses, such as $H_0 : d = 0$ and $H_1 : d = 0.2$. Originally developed in connection with Z-tests, the general framework of the SPRT can also be extended to other tests (e.g., t-tests; Hajnal, 1961; Rushton, 1950, 1952), and a similar procedure for testing mean differences has been developed within a Bayesian framework (Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, 2017).

As is indicated in Table 1, both SPRT-like procedures and our independent

segments procedure have the advantage that they can be used with virtually any test statistic. Within SPRT, the likelihood ratio $\mathbf{L}_i$ associated with H_0 versus the specific H_1 is updated after each single observation $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_i$, so the procedure can be used with any test for which this likelihood ratio can be computed. The new procedure can also be used with any statistical test, because it is based solely on observed interim $\mathbf{p}_{obs,k}$ values under H_0, which are available in every hypothesis testing situation.

The independent segments procedure differs from SPRT procedures in that it does not require *a priori* specification of a hypothesized true effect size d under the alternative hypothesis H_1. Thus, the new procedure can be used with a composite H_1 rather than only with a simple H_1. This is an advantage when researchers simply want to determine whether a theoretically important effect is present, because in these cases it could be difficult or impossible to specify a minimum effect size in advance, making it problematic to use SPRT.

The independent segments procedure also differs from SPRT in that it requires simpler computations, at least when using standard statistical packages. Such packages routinely provide the observed p values needed when using the independent segments procedure, whereas computations of likelihood ratios would require special software that would differ for each test statistic. As Albers (2019) put it, with SPRT "the applied researcher cannot rely on easy-to-use software .... One has to work through extensive technical textbooks in order to use this method" (p. 3).

Finally, the new approach also differs from SPRT procedures with respect to required sample sizes. Although the SPRT requires the smallest sample sizes *on average*, the required sample size for any individual study is unpredictable, and in fact has no theoretical upper limit. Uncertainty about sample size complicates research planning, grant writing, and so on, because a given study's demands on the experimenter's time, lab resources, and budget cannot be determined in advance. In fact, such uncertainty about sample size may be intolerable for research projects that must be completed within limited time-spans or budgets.[1] In contrast, the maximum

[1] On the other hand, Frick (1998) argued that a fixed upper limit on sample size would be inefficient in

sample size can be prespecified within the new approach, which would be a useful practical advantage in such cases.

The second major class of sequential designs proposed previously is more similar to our current proposal. These group-sequential (GS) designs (e.g., Armitage, McPherson, & Rowe, 1969; Pocock, 1977; for reviews, see Proschan et al., 2006; Wassmer & Brannath, 2016) are similar to the independent segments approach in that a maximum sample size can be specified in advance and interim data are checked at certain points before the full sample is collected. Various criteria have been suggested for using the interim data to decide how to proceed after each check. A fundamental difference between the GS procedures and the independent segments approach, however, is that GS analyses are based on all accumulated data up to the check point rather than treating each new data segment independently.

Although some GS procedures also require specification of an assumed effect size[2], our new sequential approach differs from most GS procedures mainly in generality. GS procedures were developed specifically for normally-distributed test statistics (i.e., Z-test). This approach has since been extended to certain other test statistics (e.g.,

---

some situations because "the only practical response to $p = .06$ is to test more subjects" (p. 691). He proposed a different sequential procedure in which sample sizes can also be unlimited. This test has not been widely used, however, perhaps because it is mathematically intractable and computer simulations are required in order to determine appropriate stopping criteria.

[2] With some GS procedures, d must also be specified to make interim calculations of the likelihood that H_0 will be rejected once the full sample has been collected. More technically, d must be specified within GS procedures in order to allow stopping because H_0 is unlikely to be rejected (i.e., so-called "stochastic curtailment for futility"; Proschan et al., 2006, chapter 3). The hypothesized value of d is needed to determine the probability that the additional data will produce a significant result, and researchers would have to choose an arbitrary cut-off for how low that probability must be in order for sampling to be stopped. The original GS procedures did not allow early stopping for futility—only stopping when interim evidence was sufficient to reject H_0, so they did not require specification of d. These procedures were unsatisfactory, however, because researchers were required to continue sampling even when initial results suggested that no effect would be found. This would be especially wasteful of study resources when the base rate of true effects was low.

two-sample t-statistics & some nonparametric tests; Lin, 1991; Proschan et al., 2006), but this must be done for each statistic separately, so researchers may find that a needed extension is not available. Moreover, even when the extensions are available, at each step the GS approach requires calculations provided only by specialized statistical packages (Zhu, Ni, & Yao, 2011). By contrast, since the new independent segments procedure requires only observed $\mathbf{p}_{obs,k}$ values, it is general enough to be used easily with any test statistic.

Given that the independent segments procedure has several attractive features in comparison with GS procedures, it is reasonable to ask how it compares with respect to sample sizes for equal levels of α and power. Figure 3 shows how the expected sample sizes, $\mathrm{E}[\mathbf{N}]$, of the independent segments procedure compare with those of two standard GS procedures and the fixed-sample procedure. The GS procedures are those of Pocock (1977) and O'Brien and Fleming (1979), elaborated to allow early stopping for futility using the method of DeMets and Ware (1980)[3]. Comparisons are shown for different levels of α, power $1-\beta$, and effect size d. Computations are also shown for different values of the base rate of the true effects, π, since researchers must use the same experimental procedure whether a true effect is present or not and since the base rate of true effects is generally unknown and probably differs among research areas (e.g., Wilson & Wixted, 2018). Computations for all procedures used one-sample Z-tests. $k_{max} = 3$ segments were used for the independent segments procedure, and for comparability three checks were also used for the GS procedures. For each procedure, the first step was to compute the per-segment or per-check sample size n_s needed to achieve the indicated power $1-\beta$ for the indicated effect size d.[4] Then, the overall expected sample size $\mathrm{E}[\mathbf{N}]$ was computed assuming that this sample size was used in all

[3] The constants associated with early stopping for futility for these procedures were taken from Table 2.9 of Wassmer and Brannath (2016). They recommend using the constants corresponding to Z values of either 0 or -0.5, and we arbitrarily chose to show the results for constants associated with $Z = 0$. Very similar results are obtained using the values associated with $Z = -0.5$ instead.

[4] For maximal precision, these computations allowed sample sizes to be real numbers rather than whole numbers, which would be required in practice.

studies and that the true effect of d was present in a proportion π of all studies, with H_0 being true in the remaining $1 - \pi$ of the studies.

Naturally, as can be seen by comparing across panels, the expected sample size increases with the required power level for all procedures, and likewise larger samples are needed to achieve a given power with smaller true effects. Most importantly for the present purposes, however, there are only slight differences between the independent segments procedure and the GS procedures. Thus, the less general and more complicated GS procedures do not seem to be greatly superior even when all of their assumptions are met. In fact, the independent segments procedure actually seems slightly better than these procedures when the base rate of true effects is less than approximately 0.20. Interestingly, expected sample sizes also tend to increase with the base rate of true effects, evidently because the chances of stopping early (for futility) are lower when true effects are more common.

In practice, of course, researchers must choose sample sizes without knowing the true effect size, so it is also appropriate to compare the efficiencies of the different procedures with unexpected effect sizes. Figure 4 provides such a comparison, again for the 1-sample Z test. For each procedure, we first identified the fixed or segment sample size required to give power $1 - \beta = 0.8$ for an effect size of $d = 0.5$. Then, we computed the expected sample size and power for the procedure using this same identified sample size for each value of d. Most importantly, the different sequential procedures all show similar sample size reductions relative to the fixed-sample procedure, regardless of the actual effect size. All of the sequential procedures are attractive because they use larger samples with true effects of intermediate sizes, whereas they use smaller samples with large effects that are easier to establish and with small effects that are usually of less interest. The independent segments procedure seems to use slightly smaller samples than the other two sequential procedures for small effect sizes, whereas the Pocock procedure does slightly better for large ones. Importantly, the different sequential procedures all have power that is virtually identical to that of the fixed-sample procedure for all effect sizes—not just for the effect size of $d = 0.5$ for which power was

equated.

General Discussion

We have proposed a new independent segments strategy for testing directional hypotheses that is an alternative both to the fixed-sample strategy that is the de facto standard in many research areas and to traditional sequential strategies. Like these other strategies, the new approach results in familiar dichotomous decisions that are needed for rational all-or-none decision making (e.g., "are the results sufficiently strong to recommend a drug or treatment?"). Like other sequential strategies, the new procedure is more efficient than the fixed-sample strategy in that it yields equally accurate decisions with smaller on-average sample sizes. Compared to other sequential strategies, the new strategy is relatively simple because it requires only consideration of $\mathbf{p}_{obs,k}$ values of any statistical tests (e.g., t-tests).

Choice of k_{max}, α_{strong}, and n_s

For researchers who wish to adopt the independent segments approach, a natural question is how to choose the best values of k_{max}, α_{strong}, and n_s. Fortunately, because of the mathematical simplicity of the independent segments procedure this optimization process is straightforward in principle, though it is computationally intensive. As is described in Appendix B, R software is provided to facilitate the process. We will illustrate the process in this section by way of an example.

Assume that a researcher intends to conduct a study using a 2-sample t-test and wants an overall Type I error rate of $\alpha = 0.05$. Assume further that the researcher would like to have a power level of $1 - \beta = 0.8$ to detect an effect of $d = 0.5$. Based on these assumed specifications, a fixed-sample researcher would compute that a sample size of 100 was required.

To find the best values of k_{max}, α_{strong}, and n_s, the independent segments researcher must use numerical search. Given the requirements of $\alpha = 0.05$ and $1 - \beta = 0.8$ for an effect of $d = 0.5$, the best parameters for the independent segments researcher are simply those parameters that minimize the expected sample size, $\mathrm{E}[\mathbf{N}]$.

The numerical search routine is set up to search the parameter space defined by different combinations of k_{max} and α_{strong}. For each combination, it is possible to compute—again by numerical search—the per-segment sample size, n_s, that is required to attain the desired power with the desired effect and α level (see Appendix A for details). For example, Figure 5A shows the required n_s values for this example for many combinations of k_{max} and α_{strong}.

Once the required n_s has been determined for a given combination of k_{max} and α_{strong}, the next step is to compute the expected total study sample size, $\mathrm{E}[\mathbf{N}]$, for that combination. Because the expected sample size depends on the true effect size (e.g., Figure 4), however, it is necessary at this point to consider a full hypothetical distribution of possible true d values. For example, Figure 5B shows the expected sample size when the true effect is always the target value of $d = 0.5$. In contrast, Figure 5C shows the averages of the expected sample sizes with $d = 0$ and $d = 0.5$, which correspond to the overall expected sample size for a research scenario in which $d = 0$ and $d = 0.5$ are equally likely. Figure 5D shows the overall expected sample size under the scenario that the true d values have a uniform distribution between zero and one. Fortunately, the best values of k_{max} and α_{strong}—that is, those yielding the minimum expected sample size—are similar for all of these cases. Specifically, a good choice for this example appears to be $k_{max} = 4$ and $\alpha_{strong} \approx 0.035$, for which the required n_s would be approximately 48, as shown in Figure 5A. More generally, researchers with more information about the possible d values could make an analog of Figure 5D for any desired mixture distribution of d values. Not surprisingly, the optimal choices of k_{max} and α_{strong} (i.e., for minimizing expected sample size) depend on the experimental design and on the hypothesized distribution of possible true d values, but it is a nice feature of the independent segments procedure that optimal settings can be determined by direct calculation for any specific assumed research scenario.

Practical Limitations

Several practical considerations may limit the usefulness of the independent segments strategy suggested here. One is that the strategy is only directly applicable when a single key hypothesis is tested in each study. As is true for all sequential procedures, deciding when to stop is more complicated when multiple hypotheses are tested in the same study (e.g., in factorial designs), because then preliminary evidence about different hypotheses must be combined to make a single overall decision about stopping. It should be emphasized, however, that the single key hypothesis need not concern a main effect, although it does in our examples. Instead, a researcher might be primarily interested in detecting correlations, interactions, odds ratios greater than one, or almost any other statistical pattern for which a $\mathbf{p}_{obs,k}$ value can be computed.

Likewise, the present strategies would not be useful for studies designed to produce evidence for the absence of an effect—that is, to show that H_0 is not far wrong. Indeed, NHST is generally inappropriate for this purpose (e.g., Amrhein, Greenland, & McShane, 2019). This is probably not a serious limitation in practice, because most researchers do seek to show that hypothesized effects are present (e.g., Fanelli, 2012; Leggett, Thomas, Loetscher, & Nicholls, 2013; Masicampo & Lalande, 2012). When the goal is to support models predicting that an effect should be absent, however, it is clearly necessary to obtain a precise estimate of the true effect size, and this can only be done by collecting many observations—not by stopping early.

In light of recent concerns about the extent of "p-hacking" in data analysis (e.g., Head, Holman, Lanfear, Kahn, & Jennions, 2015), it seems important to ask what effect the independent segments strategy might have on researchers' tendencies to perform questionable analyses. The answer is surely very complicated, because it depends on researchers' understanding of and attitudes toward proper analysis procedures. Two consequences of the independent segments strategy seem particularly salient in this context, and they seem to work in opposite directions with respect to the number of true versus false positives (TPs and FPs) in the literature.

First, early FTR decisions reduce researchers' sunk costs in a given H_0 test,

thereby reducing the researchers' motivation to perform questionable analyses in order to find *something* publishable after collecting a large data set. This feature of the independent segments procedure would tend to reduce the number of FPs in the literature.

Second, because of its increased efficiency, the independent segments approach would allow researchers to conduct more studies, and this would increase the number of FPs—as well as TPs—in the literature. In essence, using the independent segments procedure is equivalent to increasing the research budget because this procedure allows more studies to be carried out with no reduction in power. The relative proportions of TPs versus FPs would not be altered by using the independent segments procedure, as long as current α and power levels were maintained. The separate question of what α and power levels are optimal in different research areas can only be examined within the context of an explicit model for the costs and benefits associated with different study outcomes (e.g., Miller & Ulrich, 2016, 2019).

A final set of practical issues concerns publication of research using the independent segments strategy. For a full disclosure of the research protocol, the exact hypothesis testing strategy must of course be stated in the methods section. We would anticipate that reviewers and editors would basically accept the new strategy as equivalent to the fixed-sample one, since the Type I error probability can be set to any required level (e.g., $\alpha = 0.05$). In describing the results, it seems appropriate not only to mention the exact series of $\mathbf{p}_{obs,k}$ values, but also to report analyses of the combined results across all segments, both for simplicity and for maximum comparability with previous descriptions of fixed-sample analyses. It must be kept in mind, however, that an effect judged significant by the independent segments procedure need not reach significance in the combined analysis. Furthermore, the observed size of significant effects is likely to be larger than the true effect size when H_0 is rejected based on $\mathbf{p}_{obs,k} < \alpha_{strong}$, because of the bias inherent in both fixed-sample (e.g., Ulrich, Miller, & Erdfelder, 2018) and sequential (e.g., Govindarajulu, 2004; Pinheiro & DeMets, 1997; Proschan et al., 2006) hypothesis testing procedures. It has thus been important to

develop methods for obtaining less-biased effect-size estimates for both fixed-sample and group-sequential designs (e.g., Fan, DeMets, & Lan, 2004; Ulrich et al., 2018), and future research on this topic would also be needed for the independent segments procedure.

Conclusions

The present results indicate that independent segments hypothesis testing can substantially increase research efficiency. Similar to other sequential procedures, it requires smaller samples, on average, to achieve the same levels of statistical power as fixed-sample testing. The new procedure should thus be considered as a viable alternative to previous sequential procedures, and it could reasonably be the procedure of choice based on its simplicity and generality.

Table 1

Feature comparison of sequential probability ratio test (SPRT), group-sequential (GS) test, and independent segments procedure.

	Testing Procedure		
Test Property	SPRT	GS	Segments
Usable with any test statistic	Yes	No	Yes
Usable without assuming an effect size d	No	Some	Yes
Simple calculations	No	Some	Yes
Maximum sample size known in advance	No	Some	Yes

Figure 1. Illustration of the independent segments procedure for sequential 1-tailed testing of a null hypothesis (H_0). The diamonds depict a series of at most $k_{max} = 3$ independent data collection segments, with n_s observations per segment. To achieve a specific overall Type 1 error rate of α, the procedure uses two cutoff values, α_{strong} and α_{weak}, with $\alpha_{strong} < \alpha < \alpha_{weak}$, whose computation is explained in Appendix A. The first segment's data are used to test H_0 with a standard statistical method (e.g., t-test), and the resulting $\mathbf{p}_{obs,k}$ determines what happens at the end of segment 1: (a) If $\mathbf{p}_{obs,k} \leq \alpha_{strong}$, the study is terminated with the decision of rejecting H_0. (b) If $\mathbf{p}_{obs,k} > \alpha_{weak}$, the study is terminated with the decision of failing to reject H_0 (FTR). (c) If $\alpha_{strong} < \mathbf{p}_{obs,k} \leq \alpha_{weak}$, another segment of data is collected. If data are collected for further segments, the same rules are applied using the new values of $\mathbf{p}_{obs,k}$ computed from the data of each new segment considered in isolation. If data are collected for the final segment k_{max} (here, 3), the study always terminates and only two outcomes are possible depending on the $\mathbf{p}_{obs,k}$ value for that segment: (a) If $\mathbf{p}_{obs,k} \leq \alpha_{weak}$, H_0 is rejected. (b) If $\mathbf{p}_{obs,k} > \alpha_{weak}$, H_0 is not rejected. Appendix A also shows how to compute power and expected sample size for this procedure.

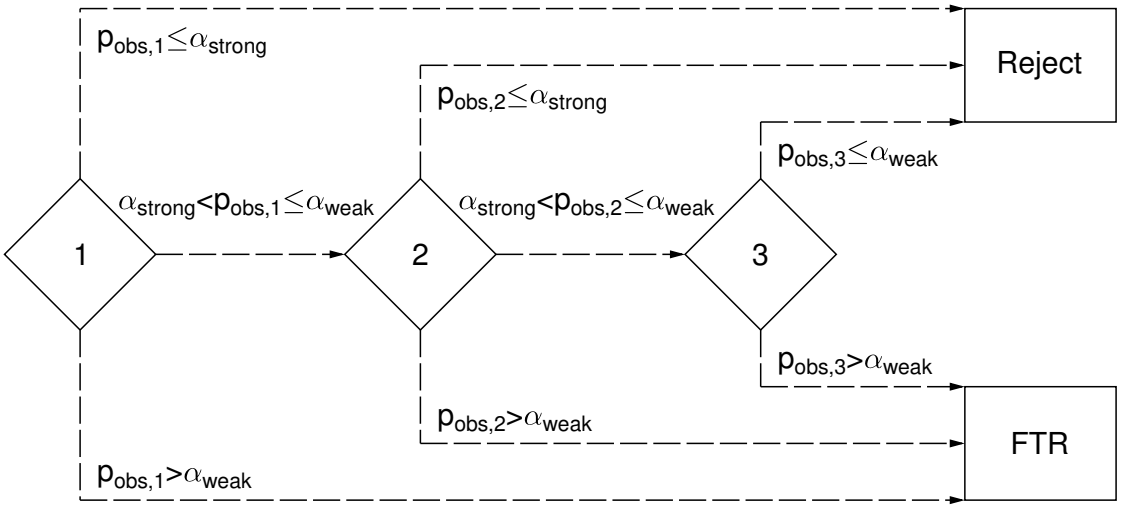
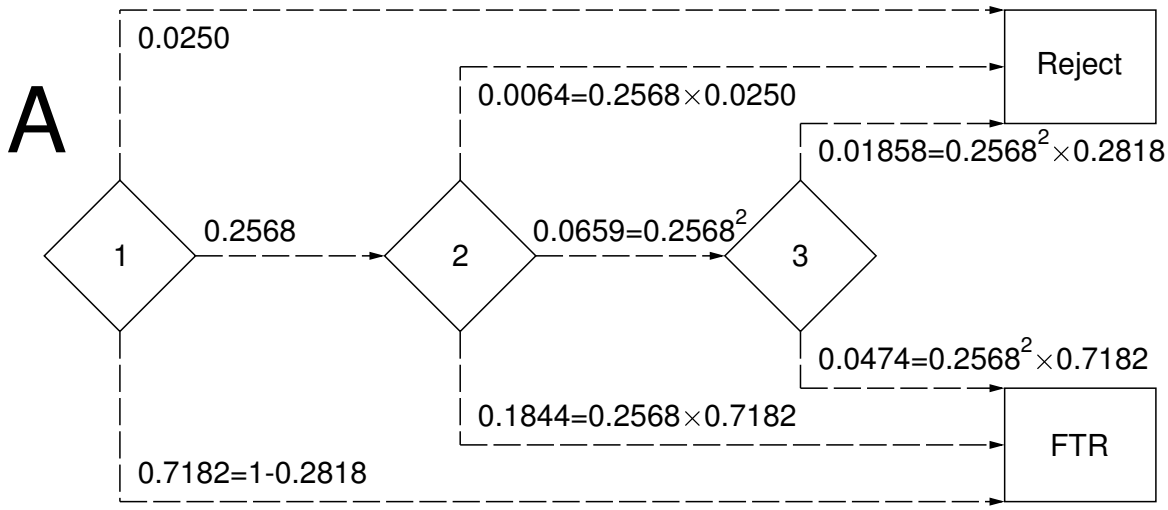


Figure 2. A numerical illustration of the probabilities associated with the different transitions for a researcher using the three-segment strategy shown in Figure 1. This researcher uses $\alpha_{weak} = 0.2818$ and $\alpha_{strong} = 0.025$, which together produce an overall Type I error rate of $\alpha = 0.05$. The researcher conducts 2-sample t-tests with a sample size of 50 per segment (i.e., 25 per group per segment). A: Probabilities for the case in which the researcher tests a true H_0 (i.e., $d = 0$). B: Probabilities for the case in which the researcher tests a false H_0 with $d = 0.5$. d is the true difference in group means divided by the common true standard deviation.



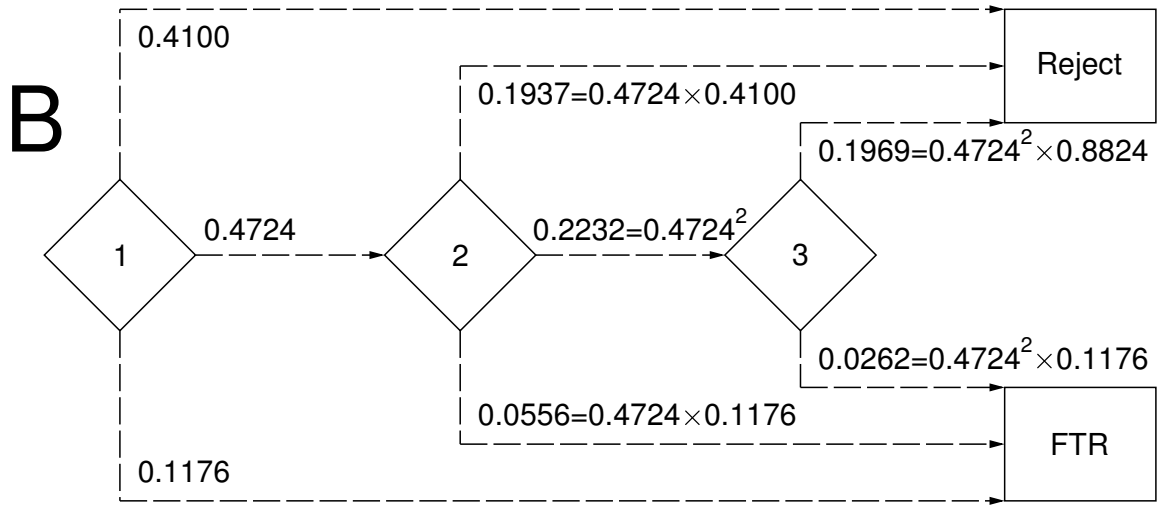


Figure 3. Expected sample size, E [**N**], for hypothesis testing with the fixed-sample procedure, the independent segments procedure, and group-sequential procedures based on the procedures of Pocock (1977) and O'Brien and Fleming (1979). E [**N**] is shown as a function of the base rate of true effects (π, abscissa), the size of the true effect when it is present d, and experimental power, $1 - \beta$. Computations used one-sample Z-tests with $\alpha = 0.025$, one-tailed. Computations for the independent segments procedure used $k_{max} = 3$ segments and $\alpha_{strong} = 0.01$. Computations for the group-sequential procedures also used three data checks with the criterion of $Z < 0$ for stochastic curtailment due to futility.

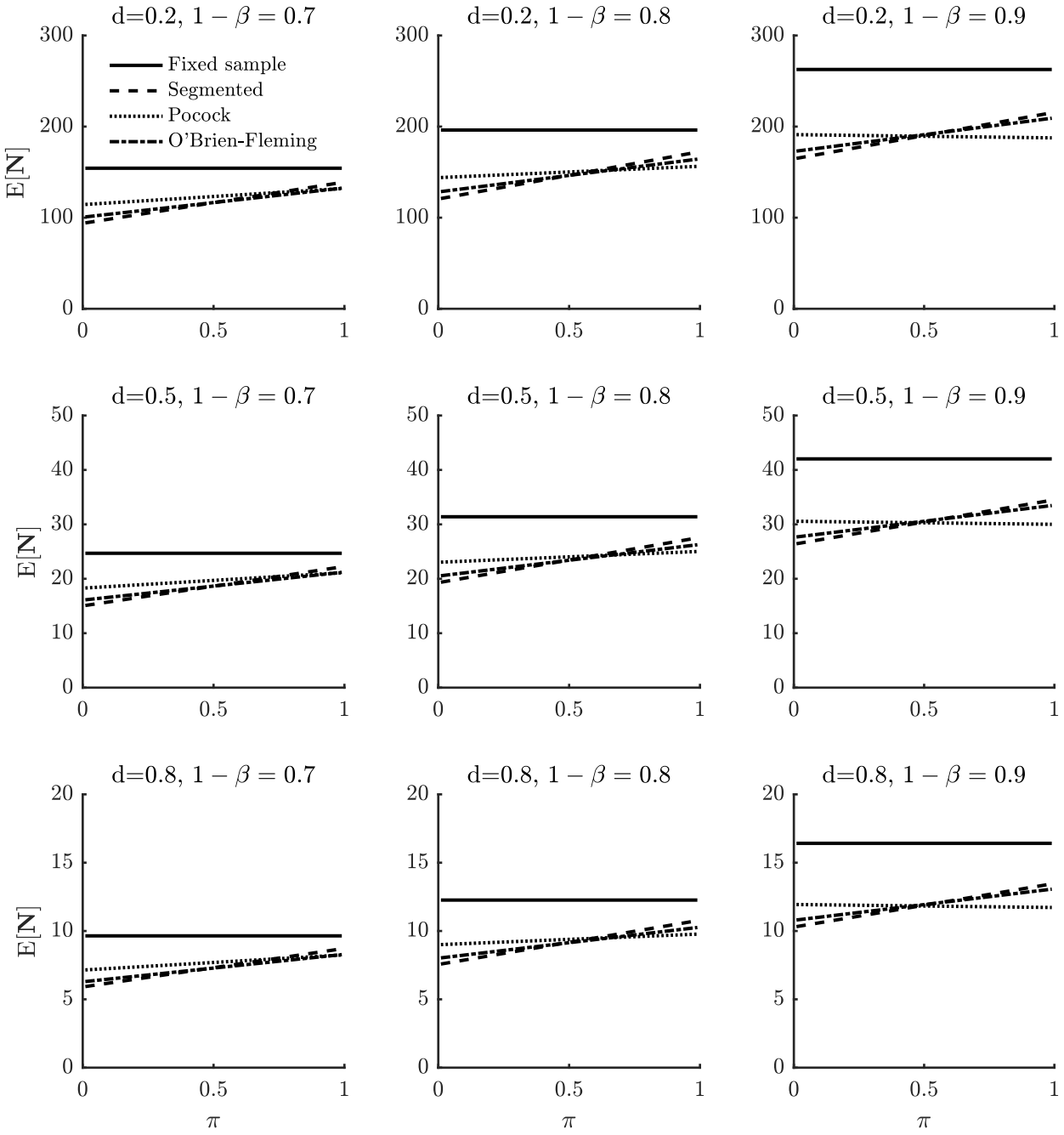


Figure 4. Power, $1 - \beta$, and expected sample size, $\mathrm{E}[\mathbf{N}]$, for hypothesis testing with the fixed-sample procedure, the independent segments procedure, and group-sequential procedures based on the procedures of Pocock (1977) and O'Brien and Fleming (1979). The sample size was $N = 31.40$ for the fixed-sample procedure, and the segment sizes were $n_s = 15.03$, 12.49, and 10.98 for the independent segments procedure, Pocock, and O'Brien-Fleming procedures, respectively. The other parameters were the same as those used in Figure 3.

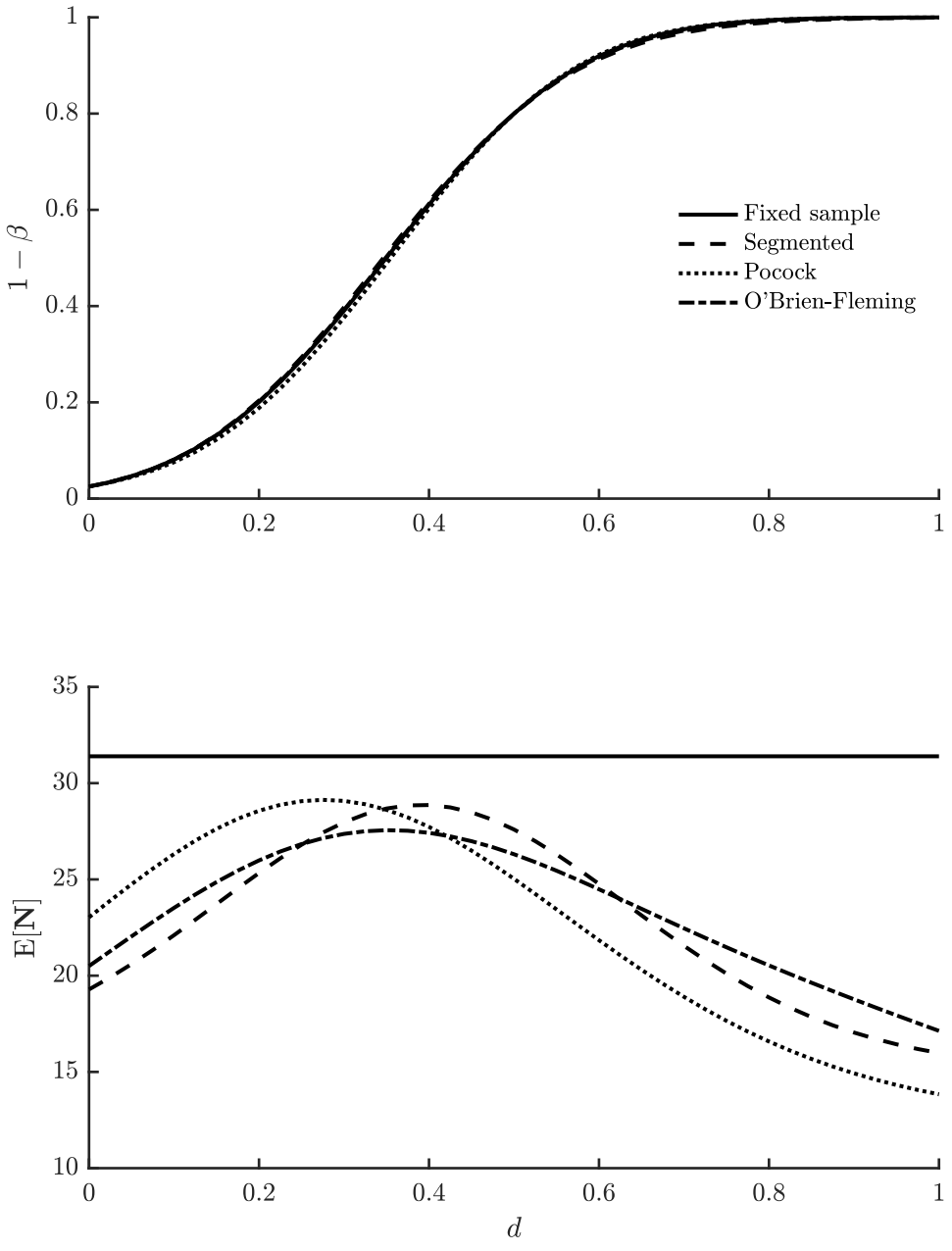
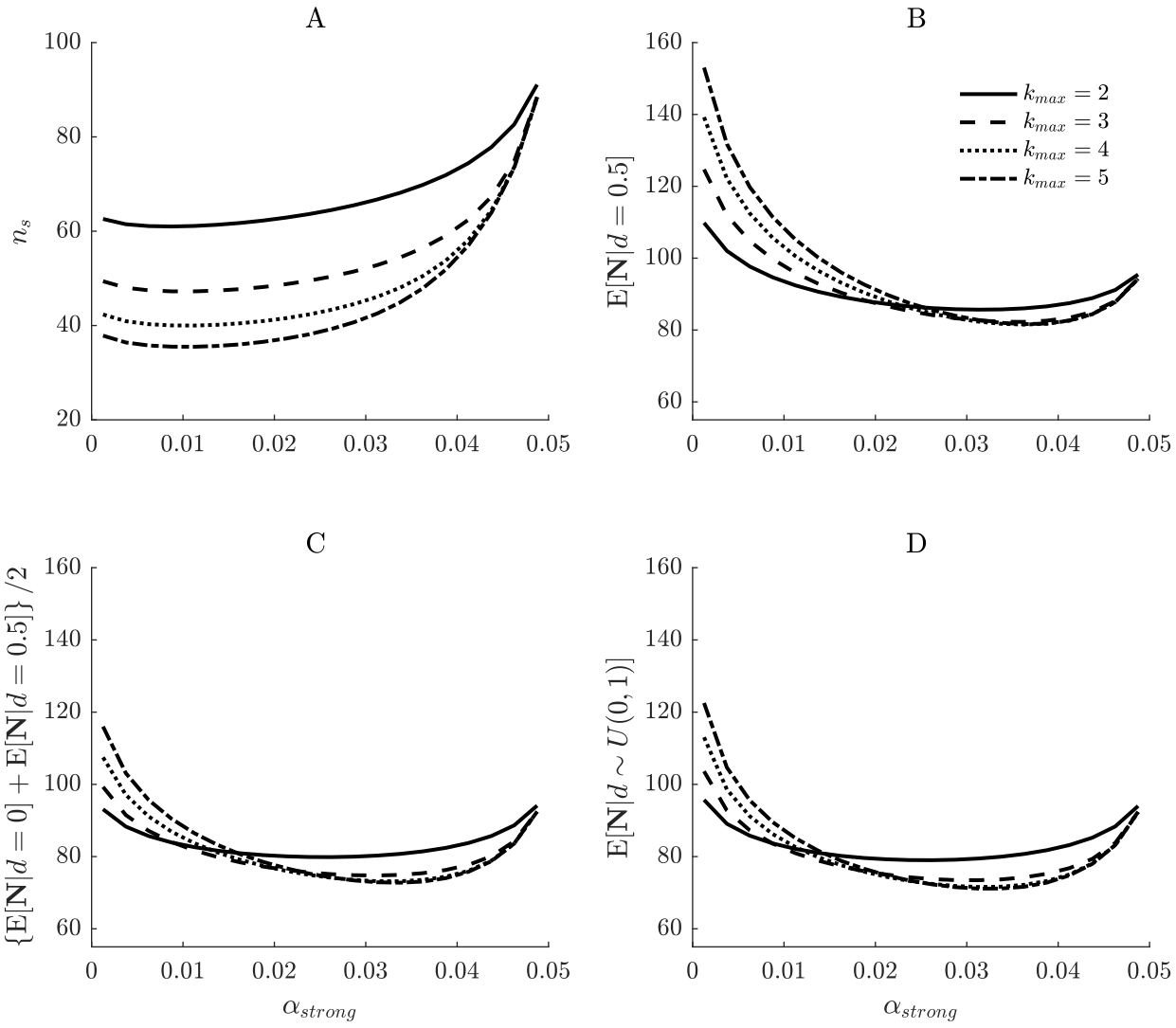


Figure 5. Illustration of computations needed to choose the optimal values of k_{max}, α_{strong}, and n_s for the example in the text. A: The value of n_s needed to achieve the desired power level of $1 - \beta = 0.8$ for the effect size of $d = 0.5$. B: The expected study sample size when $d = 0.5$, $\mathrm{E}[\mathbf{N}]$, for the indicated combination of k_{max} and α_{strong} with the associated sample size n_s shown in A. C: The average expected study sample size, $\mathrm{E}[\mathbf{N}]$, averaging across $d = 0$ and $d = 0.5$, for the indicated combination of k_{max} and α_{strong} and testing done with the associated sample size n_s shown in A. D: Analogous to panel C, but averaging $\mathrm{E}[\mathbf{N}|d]$ across d values distributed uniformly from 0–1.



Appendix A

Formal Analysis of the Independent Segments Procedure

This appendix presents a formal analysis of the independent segments procedure (ISP) diagrammed in Figure 1. To use this strategy, the researcher must specify three parameters: k_{max}, the maximum number of segments used to test H_0; α_{weak}, the significance cut-off that must be attained in each segment in order to continue the series of segments; and α_{strong}, a cut-off value such that the researcher immediately stops and rejects H_0 if $\mathbf{p}_{obs,k} \leq \alpha_{strong}$ in any segment. The values of these three parameters can be chosen to obtain any desired overall α level for the study as a whole. Once these parameters have been selected, researchers can then compute the sample size per segment, n_s, that is required to attain any desired level of statistical power for any given effect size.

With this strategy there are three possible decisions after each of the first $(k_{max} - 1)$ segments, with the following probabilities:

$$\Pr\left(\text{decision}|\mathbf{K} < k_{max}\right) = \begin{cases} \text{reject} & \Pr\left(\mathbf{p}_{obs,k} \leq \alpha_{strong}\right) \\ \text{FTR} & \Pr\left(\mathbf{p}_{obs,k} > \alpha_{weak}\right) \\ \text{continue} & \Pr\left(\alpha_{strong} < \mathbf{p}_{obs,k} \leq \alpha_{weak}\right). \end{cases}$$

At the final segment, there are only two possible decisions:

$$\Pr\left(\text{decision}|\mathbf{K} = k_{max}\right) = \begin{cases} \text{reject} & \Pr\left(\mathbf{p}_{obs,k} \leq \alpha_{weak}\right) \\ \text{FTR} & \Pr\left(\mathbf{p}_{obs,k} > \alpha_{weak}\right). \end{cases}$$

The probability of stopping after $\mathbf{K}$ segments and rejecting H_0 is thus:

$$\Pr\left(\mathbf{K} = k \;\&\; \text{decision} = \text{reject}\right) = \begin{cases} \Pr\left(\mathbf{p}_{obs,k} \leq \alpha_{strong}\right) \times \Pr\left(\alpha_{strong} < \mathbf{p}_{obs,k} \leq \alpha_{weak}\right)^{k-1} \\ \qquad \text{for } k = 1 \ldots (k_{max} - 1), \\ \Pr\left(\mathbf{p}_{obs,k} \leq \alpha_{weak}\right) \times \Pr\left(\alpha_{strong} < \mathbf{p}_{obs,k} \leq \alpha_{weak}\right)^{k-1} \\ \qquad \text{for } k = k_{max}. \end{cases}$$

The total probability of rejecting H_0 (i.e., at any segment) is then the sum of the

rejection probabilities across all segments,

$$\begin{aligned}\Pr\left(\text{decision} = \text{reject}\right) &= \sum_{k=1}^{k_{max}} \Pr\left(\mathbf{K} = k \;\&\; \text{decision} = \text{reject}\right) \\ &= \sum_{k=1}^{k_{max}-1} \Pr\left(\mathbf{p}_{obs,k} \leq \alpha_{strong}\right) \times \Pr\left(\alpha_{strong} < \mathbf{p}_{obs,k} \leq \alpha_{weak}\right)^{k-1} \\ &\quad + \Pr\left(\alpha_{strong} < \mathbf{p}_{obs,k} \leq \alpha_{weak}\right)^{k_{max}-1} \times \Pr\left(\mathbf{p}_{obs,k} \leq \alpha_{weak}\right).\end{aligned}$$

Given that $\sum_{k=1}^{n-1} a \cdot b^{k-1} = a\frac{1-b^{n-1}}{1-b}$, the preceding equation can be written as

$$\begin{aligned}\Pr\left(\text{decision} = \text{reject}\right) &= \Pr\left(\mathbf{p}_{obs,k} \leq \alpha_{strong}\right) \frac{1 - \Pr\left(\alpha_{strong} < \mathbf{p}_{obs,k} \leq \alpha_{weak}\right)^{k_{max}-1}}{1 - \Pr\left(\alpha_{strong} < \mathbf{p}_{obs,k} \leq \alpha_{weak}\right)} \\ &\quad + \Pr\left(\mathbf{p}_{obs,k} \leq \alpha_{weak}\right) \times \Pr\left(\alpha_{strong} < \mathbf{p}_{obs,k} \leq \alpha_{weak}\right)^{k_{max}-1}.\end{aligned}$$

When a true effect is actually present (i.e., H_0 is false), this sum is the global power of the hypothesis testing procedure.

Correspondingly, the global Type I error rate (α) of the procedure is

$$\alpha = \alpha_{strong}\frac{1 - (\alpha_{weak} - \alpha_{strong})^{k_{max}-1}}{1 - (\alpha_{weak} - \alpha_{strong})} + \alpha_{weak}(\alpha_{weak} - \alpha_{strong})^{k_{max}-1}.$$

Thus, for any desired overall α level, maximum number of segments (k_{max}), and early rejection cut-off (α_{strong}), the appropriate value of α_{weak} can be obtained numerically as the solution of this equation. The following two sections show R and MATLAB functions to obtain such solutions. For example, to produce $\alpha = 0.05$ with $k_{max} = 3$, the researcher could use a combination of ($\alpha_{strong}, \alpha_{weak}$) such as (0.001,0.36569), (0.005,0.35437), (0.01,0.33905), or (0.025,0.28178). Example sets of admissible pairs ($\alpha_{strong}, \alpha_{weak}$) producing a desired α level for various values of k_{max} are given in Table A1. In the limiting case of $\alpha_{strong} = 0$, the corresponding value of α_{weak} is $\alpha^{1/k_{max}}$.

As a numerical example, suppose that a researcher intends to conduct a series of at most three segments, plans to stop early and reject H_0 if any segment yields $\mathbf{p}_{obs,k} \leq \alpha_{strong} = 0.01$, and wants to arrange a global Type I error rate of $\alpha = 0.05$. Table A1 indicates that the appropriate value of α_{weak} for this situation is 0.339, meaning that the H_0 will be rejected only if the effect is significant at the level of at least $\mathbf{p}_{obs,k} \leq 0.339$ in each of three segments or at the level of $\mathbf{p}_{obs,k} \leq 0.01$ in a single segment. For a researcher carrying out 2-sample t-tests with a sample size of 10 per

group in each segment, the appropriate upper t cut-off for stopping early and rejecting H_0 is $t_{strong} = F^{-1}(0.99|df = 18) = 2.55$, where F is the cumulative distribution function of the central t distribution with the indicated degrees of freedom. Correspondingly, the appropriate lower t cut-off for stopping data collection with the FTR decision is $t_{weak} = F^{-1}(1 - 0.339|df = 18) = 0.42$.

The power of the strategy can be computed similarly using the noncentral t distribution. When the true size of the effect being tested is d, the relevant probabilities can be computed from the CDF of the noncentral t distribution, $G\left[t|df, \gamma\left(d, n_s\right)\right]$, with the indicated degrees of freedom and noncentrality parameter $\gamma\left(d, n_s\right)$, which depends on d and on the sample size per segment n_s. Specifically,

$$\begin{aligned}
\Pr\left(\mathbf{p}_{obs,k} \leq \alpha_{strong}\right) &= 1 - G\left[t_{strong}|df, \gamma\left(d, n_s\right)\right] \\
\Pr\left(\alpha_{strong} < \mathbf{p}_{obs,k} \leq \alpha_{weak}\right) &= G\left[t_{strong}|df, \gamma\left(d, n_s\right)\right] - G\left[t_{weak}|df, \gamma\left(d, n_s\right)\right] \\
\Pr\left(\mathbf{p}_{obs,k} > \alpha_{weak}\right) &= G\left[t_{weak}|df, \gamma\left(d, n_s\right)\right].
\end{aligned}$$

The total sample size used to test each H_0, $\mathbf{N}$, is a random variable determined by the number of segments that are carried out. The distribution of the number of segments, $\mathbf{K}$, is

$$\Pr\left(\mathbf{K} = k\right) = \begin{cases} \Pr\left(\alpha_{strong} < \mathbf{p}_{obs,k} \leq \alpha_{weak}\right)^{k-1} \times & \\ \qquad \left[\Pr\left(\mathbf{p}_{obs,k} \leq \alpha_{strong}\right) + \Pr\left(\mathbf{p}_{obs,k} > \alpha_{weak}\right)\right] & k = 1 \ldots (k_{max} - 1), \\ \Pr\left(\alpha_{strong} < \mathbf{p}_{obs,k} \leq \alpha_{weak}\right)^{k-1} & k = k_{max} \end{cases}$$

with mean and variance

$$\begin{aligned}
\mathrm{E}\left[\mathbf{K}\right] &= \frac{1 - c^{k_{max}}}{1 - c} \\
\mathrm{Var}\left[\mathbf{K}\right] &= \frac{c + c^{k_{max}}\left[1 - c^{k_{max}} - c - 2k_{max}\left(1 - c\right)\right]}{\left(1 - c\right)^2},
\end{aligned}$$

where $c = \Pr\left(\alpha_{strong} < \mathbf{p}_{obs,k} \leq \alpha_{weak}\right)$. The expected sample size per study is then $\mathrm{E}\left[\mathbf{N}\right] = n_s \cdot \mathrm{E}\left[\mathbf{K}\right]$, where n_s is the sample size in each segment. Of course, the distribution of $\mathbf{K}$—and thus its expectation—would depend on whether H_0 was true. With a large true effect d, for example, $\Pr\left(\mathbf{p}_{obs,k} \leq \alpha_{strong}\right)$ would approach 1 and H_0 would usually be rejected after only a single segment (i.e., $\mathbf{K} = 1$).

Table A1

Combinations of α_{weak} and α_{strong} producing desired overall α levels with the independent segments procedure.

		Value of α_{strong}										
k_{max}	α	0.000	0.001	0.002	0.003	0.004	0.005	0.006	0.007	0.008	0.009	0.010
2	0.005	0.071	0.064	0.056	0.046	0.034						
2	0.010	0.100	0.095	0.090	0.085	0.079	0.073	0.066	0.058	0.049	0.036	
2	0.020	0.141	0.138	0.135	0.132	0.129	0.125	0.121	0.118	0.114	0.109	0.105
2	0.050	0.224	0.222	0.220	0.218	0.216	0.215	0.213	0.211	0.209	0.207	0.205
3	0.005	0.171	0.157	0.141	0.120	0.090						
3	0.010	0.215	0.207	0.198	0.188	0.177	0.165	0.150	0.133	0.111	0.078	
3	0.020	0.271	0.266	0.261	0.255	0.249	0.243	0.237	0.230	0.223	0.215	0.207
3	0.050	0.368	0.366	0.363	0.360	0.357	0.354	0.351	0.348	0.345	0.342	0.339
4	0.005	0.266	0.247	0.224	0.193	0.142						
4	0.010	0.316	0.305	0.293	0.280	0.265	0.247	0.226	0.199	0.161	0.102	
4	0.020	0.376	0.370	0.363	0.355	0.348	0.340	0.331	0.322	0.312	0.302	0.290
4	0.050	0.473	0.470	0.466	0.463	0.459	0.456	0.452	0.448	0.445	0.441	0.437
5	0.005	0.347	0.324	0.295	0.254	0.178						
5	0.010	0.398	0.385	0.371	0.355	0.336	0.313	0.285	0.247	0.191	0.108	
5	0.020	0.457	0.450	0.442	0.434	0.425	0.415	0.405	0.394	0.382	0.369	0.354
5	0.050	0.549	0.546	0.542	0.538	0.534	0.530	0.526	0.522	0.517	0.513	0.508
6	0.005	0.414	0.388	0.354	0.302	0.196						
6	0.010	0.464	0.450	0.434	0.415	0.393	0.366	0.329	0.278	0.204	0.109	
6	0.020	0.521	0.513	0.504	0.495	0.485	0.474	0.463	0.450	0.436	0.420	0.402
6	0.050	0.607	0.603	0.599	0.595	0.591	0.586	0.582	0.577	0.572	0.567	0.562

Note. Each tabled α_{weak} value is appropriate for the indicated number of segments (k_{max}), global Type I error rate (α), and cut-off for an early H_0 rejection (α_{strong}). Missing values indicate parameter combinations that are impossible because they violate the requirement $\alpha_{strong} \leq \alpha$.

Appendix B

Software

R functions for computations related to the Independent Segments Procedure can be downloaded from `https://github.com/milleratotago/Independent_Segments_R`. This appendix describes the main functions that would be useful for practicing researchers, as well as one equivalent MATLAB function.

R code for finding α_{weak} values

```
# Needs: library(pracma)
install.packages("pracma")
library(pracma)
find_alpha_weak <- function(overall_alpha, n_segments, alpha_strong) {
  # Use numerical search to find the appropriate alpha_weak value
  # that will produce the desired overall_alpha level for the indicated
  # values of n_segments and alpha_strong.
  compute_error <- function(try_aw) {
    diff <- try_aw - alpha_strong
    diff_accumulated <- diff ^(n_segments - 1)
    alpha_strong*(1-diff_accumulated)/(1-diff)+
      try_aw*diff_accumulated-overall_alpha
  }
  aw_range <- c(alpha_strong, overall_alpha^(1/n_segments))
  final_aw <- fzero(compute_error, aw_range)
  out <- c(final_aw$x, final_aw$fval)
  return(out)
}
# Example:
find_alpha_weak(0.05, 3, 0.01)
```

```
# should give the results 0.3390478 0.0000000
```

MATLAB code for finding α_{weak} values

```
function [aw, LastErr] = findAlphaWeak(OverallAlpha, nSegments, AlphaStrong)
    % Use numerical search to find the appropriate AlphaWeak value (aw)
    % that will produce the desired overall alpha level for the indicated
    % values of nSegments and AlphaStrong.
    awRange=[AlphaStrong, OverallAlpha^(1/nSegments)];
    [aw, LastErr] = fzero(@ComputeError,awRange);
    function err = ComputeError(tryAlphaWeak)
        Diff = tryAlphaWeak - AlphaStrong;
        DiffToPower = Diff^(nSegments-1);
        err = AlphaStrong * (1 - DiffToPower) / (1 - Diff) ...
            + DiffToPower*tryAlphaWeak - OverallAlpha;
    end
end
% Example:
[AlphaWeak, LastErr] = findAlphaWeak(0.05, 3, 0.01)
% should give the results 0.33905 0
```

Computation of Expected Outcomes

When planning a study using the ISP, it is useful to be able to check on the power and expected sample size for a given contemplated set of design parameters, and this information is provided by an R function called `segmented_hyp_test_outcomes`. For example, the R function

```
expected_summary <- segmented_hyp_test_outcomes(max_n_segments=3,
                    n_per_segment=80, alpha_total=0.05,
                    alpha_strong=0.01, stat_procedure_name='2t',
                    effect_size=0.4, base_rate=0.33)
```

reports that the design with the indicated parameters has power $1 - \beta = 0.828$ to detect the specified true effect and that the expected number of participants that will need to be tested is $\mathrm{E}[\mathbf{N}] = 130.6$. (Note that $\mathrm{E}[\mathbf{N}]$ depends on the base rate but power does not.) In addition, this function reports the probability of stopping and rejecting H_0 as well as the probability of stopping and failing to reject H_0 at each segment.

Incorporating Effect Size Targets

When researchers can specify a minimum effect size of interest a priori, this information can be used to choose optimal design parameters within the ISP, and software is provided to facilitate this process. That is, parameters can be chosen to obtain the desired level of power for detecting that effect with the minimum expected total sample size that is possible using the ISP. In this situation, though, researchers should also consider using an SPRT-based test instead if appropriate methods exist for the statistical procedure that is to be used (e.g., t-test; Schnuerch & Erdfelder, 2020).

For example, suppose the researcher wants to use a paired t-test design with overall $\alpha = 0.025$ (one tailed), $\alpha_{strong} = 0.01$, and wants to achieve power $1 - \beta = 0.80$ for a minimum effect size of $d = 0.2$ that is assumed to be present with a base rate of $\pi = 0.33$. Using the formulas in Appendix A, one can iterate over different values of the number of segments, k_{max}, to find the one yielding the minimum expected total sample size. For convenience, the R package provides the function `search_kmax` to automate this process. The R function

```
kmax_summary <- search_kmax(alpha_total=0.05, alpha_strong=0.01,
                stat_procedure_name='1t', target_power=0.8,
                effect_size=0.5, base_rate=0.33)
```

produces the output shown in Table B1, from which it can be seen that the minimum

$\mathrm{E}[\mathbf{N}]$ is achieved with $k_{max} = 3$ and $n_s = 12.26$. Researchers would naturally be forced to round n_s— presumably up to make sure of having at least the desired power—and this rounding could affect the ordering of the $\mathrm{E}[\mathbf{N}]$ values for the different k_{max} values. Thus, in a final step researchers might use the function `segmented_hyp_test_outcomes` to compare the exact $\mathrm{E}[\mathbf{N}]$ and power values for different combinations of k_{max} with integer values near the real value of n_s returned by `kmax_summary`.

Table B1

Sample output of the `search_kmax` *function*

kmax	alpha_weak	n_per_segment	exp_n_subjects	exp_n_segments
2	0.2050625	15.763986	20.79393	1.319078
3	0.3390478	12.257922	20.56697	1.677852
4	0.4366547	10.414161	21.18611	2.034356
5	0.5081760	9.392095	22.32106	2.376579
6	0.5620571	8.654113	23.41012	2.705086
7	0.6037776	8.047932	24.32973	3.023104
8	0.6368496	7.713395	25.56828	3.314790
9	0.6635800	7.354103	26.46341	3.598455
10	0.6855311	7.090773	27.42582	3.867819

Acknowledgements

We are grateful to Patricia Haden for assistance with R and MATLAB software and to Patricia Haden, Victor Mittelstädt, Edgar Erdfelder, Martin Schnuerch, and Daniel Lakens for helpful comments on earlier versions of the article. Portions of this work were presented at the 2018 annual meeting of the Psychonomic Society (Miller & Ulrich, 2018).

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

References

Albers, C. (2019, April). The problem with unadjusted multiple and sequential statistical testing. *Nature Communications*, *10*(1), 1921. Retrieved from https://doi.org/10.1038/s41467-019-09941-0 doi: 10.1038/s41467-019-09941-0

Amrhein, V., Greenland, S., & McShane, B. (2019). Retire statistical significance. *Nature*, *567*, 305–307.

Armitage, P., McPherson, C. K., & Rowe, B. C. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society: Series A (General)*, *132*(2), 235–244. doi: %2910.2307/2343787

Baker, M. (2016). Is there a reproducibility crisis? *Nature*, *533*, 452–454.

Begley, C. G., & Ioannidis, J. P. A. (2015). Reproducibility in science: Improving the standard for basic and preclinical research. *Circulation Research*, *116*(1), 116–126. doi: %8610.1161/CIRCRESAHA.114.303819

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., ... others (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433–1436.

Chalmers, I., Bracken, M. B., Djulbegovic, B., Garattini, S., Grant, J., Gülmezoglu, A. M., ... Oliver, S. (2014). How to increase value and reduce waste when research priorities are set. *Lancet*, *383*(9912), 156–165. doi: %13810.1016/S0140-6736(13)62229-1

Chalmers, I., & Glasziou, P. (2009). Avoidable waste in the production and reporting of research evidence. *Lancet*, *374*(9683), 86–89.

DeMets, D. L., & Ware, J. H. (1980). Group sequential methods for clinical trials with a one-sided hypothesis. *Biometrika*, *67*(3), 651–660. doi: %20010.1093/biomet/67.3.651

Fan, X. F., 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. doi: %25610.1081/BIP-120037195

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*(3), 891–904. doi: %28710.1007/s11192-011-0494-7

Freeman, L. P. (2017). *Rigor mortis: How sloppy science creates worthless cures, crushes hopes, and wastes billions.* Richard Harris Basic Books.

Frick, R. W. (1998). A better stopping rule for conventional statistical tests. *Behavior Research Methods, Instruments & Computers*, *30*(4), 690–697. doi: %31310.3758/BF03209488

Gorroochurn, P., Hodge, S. E., Heiman, G. A., Durner, M., & Greenberg, D. A. (2007). Non-replication of association studies: "Pseudo-failures" to replicate? *Genetics in Medicine*, *9*, 325–331. doi: %36310.1097/GIM.0b013e3180676d79

Govindarajulu, Z. (2004). *Sequential statistics.* World Scientific Publishing. doi: %39710.1142/5575

Hajnal, J. (1961). A two-sample sequential *t*-test. *Biometrika*, *48*(1/2), 65–75. doi: %42110.2307/2333131

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of *p*-hacking in science. *PLOS Biology*, *13*(3), e1002106. doi: %44610.1371/journal.pbio.1002106

Ioannidis, J. P. A. (2005). Contradicted and initially stronger effects in highly cited clinical research. *JAMA*, *294*(2), 218–228. doi: %47910.1001/jama.294.2.218

Ioannidis, J. P. A., Greenland, S., Hlatky, M. A., Khoury, M. J., Macleod, M. R., Moher, D., ... Tibshirani, R. (2014). Increasing value and reducing waste in research design, conduct, and analysis. *Lancet*, *383*(9912), 166–175. doi: %50510.1016/S0140-6736(13)62227-8

Jennison, C., & Turnbull, B. W. (2000). *Group sequential methods with applications to clinical trials.* Taylor & Francis. Retrieved from `https://books.google.co.nz/books?id=7FQwngEACAAJ`

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, *44*(7), 701–710. doi: %53710.1002/ejsp.2023

Leggett, N. C., Thomas, N. A., Loetscher, T., & Nicholls, M. E. R. (2013). The life of *p*: "Just significant" results are on the rise. *Quarterly Journal of Experimental Psychology*, *66*(12), 2303–2309. doi: %59310.1080/17470218.2013.863371

Lin, D. Y. (1991). Nonparametric sequential testing in clinical trials with incomplete multivariate observations. *Biometrika*, *78*(1), 123–131. doi: %62610.1093/biomet/78.1.123

Macleod, M. R., Michie, S., Roberts, I., Dirnagl, U., Chalmers, I., Ioannidis, J. P., ... Glasziou, P. (2014). Biomedical research: Increasing value, reducing waste. *Lancet*, *383*(9912), 101–104. doi: %65210.1016/S0140-6736(13)62329-6

Masicampo, E. J., & Lalande, D. R. (2012). A peculiar prevalence of *p* values just below .05. *Quarterly Journal of Experimental Psychology*, *65*(11), 2271–2279. doi: %68310.1080/17470218.2012.711335

Miller, J. O., & Ulrich, R. (2016). Optimizing research payoff. *Perspectives on Psychological Science*, *11*(5), 664–691. doi: %71210.1177/1745691616649170

Miller, J. O., & Ulrich, R. (2018). *Optimizing research payoff, I: A simple way to increase research efficiency.* (Presentation at the annual meeting of the Psychonomic Society, New Orleans, Nov.)

Miller, J. O., & Ulrich, R. (2019). The quest for an optimal alpha. *PLOS ONE*, *14*(1), 1–13. doi: %74110.1371/journal.pone.0208631

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., du Sert, N. P., ... Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(0021), 1–9. doi: %79710.1038/s41562-016-0021

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*(6), 615–631.

O'Brien, P. C., & Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics*, *35*(3), 549–556. Retrieved from `http://www.jstor.org/stable/2530245`

Open Science Collaboration. (2015). Estimating the reproducibility of psychological

science. *Science*, *349*(6251), aac4716-1–aac4716-8. doi: %82810.1126/science.aac4716

Pinheiro, J. C., & DeMets, D. L. (1997). Estimating and reducing bias in group sequential designs with Gaussian independent increment structure. *Biometrika*, *84*(4), 831–845. doi: %91010.1093/biomet/84.4.831

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, *64*(2), 191–199. doi: %94010.1093/biomet/64.2.191

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

Røttingen, J.-A., Regmi, S., Eide, M., Young, A. J., Viergever, R. F., Årdal, C., ... Terry, R. F. (2013). Mapping of available health research and development data: what's there, what's missing, and what role is there for a global observatory? *Lancet*, *382*(9900), 1286–1307. doi: %96610.1016/S0140-6736(13)61046-6

Rushton, S. (1950). On a sequential *t*-test. *Biometrika*, *37*(3/4), 326–333. doi: %102710.2307/2332385

Rushton, S. (1952). On a two-sided sequential *t*-test. *Biometrika*, *39*(3/4), 302–308. doi: %105210.2307/2334026

Schnuerch, M., & Erdfelder, E. (2020). Controlling decision errors with minimal costs: The sequential probability ratio *t* test. *Psychological Methods*, *25*(2), 206–226. doi: %107710.1037/met0000234

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*, *22*(2), 322–339. doi: %110610.1037/met0000061

Stevely, A., Dimairo, M., Todd, S., Julious, S. A., Nicholl, J., Hind, D., & Cooper, C. L. (2015). An investigation of the shortcomings of the CONSORT 2010 statement for the reporting of group sequential randomised controlled trials: A methodological systematic review. *PLOS ONE*, *10*(11), e0141104. doi: %113810.1371/journal.pone.0141104

Ulrich, R., Miller, J. O., & Erdfelder, E. (2018). Effect size estimation from *t*-statistics in the presence of publication bias: A brief review of existing approaches with some extensions. *Zeitschrift für Psychologie*, *226*(1), 56–80. doi: %117310.1027/2151-2604/a000319

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

Wassmer, G., & Brannath, W. (2016). *Group sequential and confirmatory adaptive designs in clinical trials.* Springer. doi: %120510.1007/978-3-319-32562-0

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

Wilson, B. M., & Wixted, J. T. (2018). The prior odds of testing a true effect in cognitive and social psychology. *Advances in Methods and Practices in Psychological Science*, *1*(2), 186–197. doi: %125610.1177/2515245918767122

Zhu, L., Ni, L., & Yao, B. (2011). Group sequential methods and software applications. *American Statistician*, *65*(2), 127–135. doi: %130910.1198/tast.2011.10213