AMERICAN PSYCHOLOGICAL ASSOCIATION

Psychological Methods

© 2021 American Psychological Association
ISSN: 1082-989X

2021, Vol. 26, No. 4, 501–506
https://doi.org/10.1037/met0000404

COMMENTARY

On the Efficiency of the Independent Segments Procedure: A Direct Comparison With Sequential Probability Ratio Tests

Edgar Erdfelder and Martin Schnuerch
Cognition and Individual Differences Lab, Department of Psychology, School of Social Sciences, University of Mannheim

Abstract

In this comment, we report a simulation study that assesses error rates and average sample sizes required to reach a statistical decision for two sequential procedures, the sequential probability ratio test (SPRT) originally proposed by Wald (1947) and the independent segments procedure (ISP) recently suggested by Miller and Ulrich (2020). Following Miller and Ulrich (2020), we use sequential one-tailed t tests as examples. In line with the optimal efficiency properties of the SPRT already proven by Wald and Wolfowitz (1948), the SPRT outperformed the ISP in terms of efficiency without compromising error probability control. The efficiency gain in terms of sample size reduction achieved with the SPRT t test relative to the ISP may be as high as 25%. We thus recommend the SPRT as a default sequential testing procedure especially for detecting small or medium hypothesized effect sizes under H_1 whenever a priori knowledge of the maximum sample size is not crucial. If a priori control of the maximum sample size is mandatory, however, the ISP is a very useful addition to the sequential testing literature.

Translational Abstract

Sequential tests analyze data sequentially when deciding between two statistical hypotheses H_0 and H_1. After each step, a decision is made whether to accept H_0, to accept H_1, or to continue sampling data, based on criteria that control error rates α (probability of accepting H_1 when H_0 holds) and β (probability of accepting H_0 when H_1 holds). Using hypotheses on means of two samples as an example, we compare the Sequential Probability Ratio Test (SPRT) originally proposed by Wald (1947) and the Independent Segments Procedure (ISP) recently proposed by Miller and Ulrich (2020). While the former method processes data cumulatively and one-by-one with the maximum sample size unknown beforehand, the latter method analyzes them independently in groups with both group size and maximum sample size known in advance. Our simulation studies show that both methods work well as they (a) keep their predefined α and β levels and (b) need smaller samples on average than the classical fixed-sample Neyman-Pearson test. However, in terms of efficiency (i.e., the average sample size required to reach a decision), the SPRT clearly outperforms the ISP (with a sample size reduction of up to 25% relative to the ISP). We conclude that the SPRT is the method of choice to minimize costs and time required for statistical decisions whenever a priori control of the maximum sample size is not necessary. If a priori control of the maximum sample size is mandatory, however, the ISP is a very useful alternative to the SPRT.

Keywords: sequential testing, sequential probability ratio tests, independent segments procedure

Miller and Ulrich (2020) proposed a new independent segments procedure (ISP) of sequential hypothesis testing that controls overall Type-1 (α) and Type-2 (β) error probabilities when deciding between a statistical null hypothesis (H_0) and an alternative hypothesis (H_1). Their ISP resembles certain group-sequential designs (cf. Lakens, 2014; Pocock, 1977; Proschan et al., 2006; Zhu et al., 2011) in that—after collecting each predefined subsample s, $s = 1, \ldots S$, of the total data set—a decision is made whether to terminate sampling and accept H_1, to terminate sampling and accept H_0, or to continue by sampling an additional subset (or segment) of predefined size n_s, provided that the maximum total sample size $N_{\max}$ has not been reached yet (i.e., $n_1 + \ldots + n_S \leq N_{\max}$). Most importantly in the present context, the ISP is innovative in that it analyzes later sampled segments independently from previous subsamples, a feature that greatly simplifies calculation of overall error probabilities. For example, when making a decision after the second segment, only the data in the second segment affect this decision whereas the data obtained in the first segment are

Edgar Erdfelder https://orcid.org/0000-0003-1032-3981
Martin Schnuerch https://orcid.org/0000-0001-6531-2265

Article preparation was supported by the Research Training Group "Statistical Modeling in Psychology," funded by the Deutsche Forschungsgemeinschaft (GRK 2277).

Correspondence concerning this article should be addressed to Edgar Erdfelder or Martin Schnuerch, Cognition and Individual Differences Lab, Department of Psychology, School of Social Sciences, University of Mannheim, A5 Building, C209, D-68131 Mannheim, Germany. Email: erdfelder@uni-mannheim.de or martin.schnuerch@psychologie.uni-mannheim.de

ignored. Intuitively, such a procedure appears inefficient as statistical decisions are made in ignorance of all data sampled from the same underlying population prior to the most recent segment. Somewhat surprisingly, however, Miller and Ulrich (2020) were able to show that their ISP compares quite well with some established group-sequential methods in terms of efficiency and in fact may even be somewhat more efficient than these methods under some conditions. Clearly, this is a nontrivial and actually very surprising result given that subsamples are analyzed independently in the ISP but cumulatively in group-sequential designs. Taking additional advantages of the ISP into account (see Table 1 in Miller & Ulrich, 2020), the ISP can be considered a very valuable innovation in the sequential testing literature that has the potential to serve as the method of choice whenever researchers engage in sequential testing.

The main purpose of our commentary is to discuss the ISP critically by extending the efficiency comparisons of the ISP to sequential probability ratio tests (SPRTs; Schnuerch & Erdfelder, 2020; Wald, 1947). Notably, Miller and Ulrich (2020, p. 3) concede that their ISP cannot be expected to compete with the SPRT in terms of efficiency when testing a point H_0 (such as $\delta = (\mu_1 - \mu_0)/\sigma = 0$ for the two-groups *t* test) against a specific point H_1 (e.g., $\delta = .50$, a "medium" effect, cf. Cohen, 1988). In contrast to both the ISP and group-sequential designs, the SPRT does not analyze the data in segments or groups. Rather, after each sampled data point, a decision is made based on the likelihood ratio of the observed data under H_1 and H_0, respectively, whether to stop sampling and accept H_1, to stop sampling and accept H_0, or to continue by sampling an additional data point. Thus, observed data are analyzed cumulatively and step-by-step in smallest possible units. Although it can be shown that the SPRT must eventually reach a decision for either H_1 or H_0 with a finite sample size (Wald, 1947, p. 157), there is no definite upper limit to the sample size known in advance. Hence, other than in group-sequential designs and in the ISP, a maximum sample size $N_{\max}$ cannot be defined a priori for the SPRT.

As already proven by Wald and Wolfowitz (1948), the SPRT is the most efficient test procedure for tests between simple hypotheses. In other words, when the true parameter corresponds to either a simple H_0 or a simple H_1, there is no alternative sequential test procedure that controls α and β error probabilities with smaller expected sample sizes. Although we have known this fact already for decades, it still is of considerable interest to compare the new ISP method with the SPRT directly. One reason is that Wald and Wolfowitz's (1948) proof does not apply to composite hypotheses that dominate behavioral research, for example, hypotheses about mean differences with unknown variance. In sequential SPRT *t* tests, therefore, the optimum property may only hold asymptotically (Lai, 1981). Another reason is that only direct comparisons can tell us how much efficiency is actually lost when using the new method instead of the SPRT. Because Miller and Ulrich (2020) do not provide such a direct comparison, we add it in this comment by reporting a Monte-Carlo study of ISP and SPRT performance.

Notably, both the ISP and SPRT are general procedures that can be adapted basically to any statistical test. However, for the sake of simplicity and brevity, we follow Miller and Ulrich (2020) and consider the one-tailed *t*-test scenario as a prototypical example only. Our simulation study and its results will be described in the next section. This is followed by a discussion of the results and some recommendations for implementing and choosing among different sequential procedures.

A Monte-Carlo Study Comparing One-Tailed SPRT and ISP *t* Tests

Method

Using a sequence of simulation studies conducted in R[1], we compared the SPRT with the ISP for 12 different *t* test simulation scenarios. Following Miller and Ulrich (2020), we employed one-tailed *t* tests between H_0: $\delta = 0$ and H_1: $\delta > 0$. Note that SPRT results for this scenario must necessarily differ from corresponding results reported in Schnuerch and Erdfelder (2020), because the latter refer to two-tailed rather than one-tailed *t* tests. Here, we simulated four population scenarios in line with the distributional assumptions of the two-groups *t* test (normal distributions within groups with homogeneous variance σ^2). One scenario conforms to H_0: $\delta = 0$, whereas the other three correspond to H_1: $\delta > 0$, with $\delta = .2$, $\delta = .5$, and $\delta = .8$ representing "small," "medium," and "large" effect sizes under H_1, respectively, according to Cohen's (1988) effect size conventions. We applied three different versions of both the SPRT and the ISP to each of these four simulated population scenarios. The three versions differed in the population effect size *d* assumed in the sequential procedures under H_1, again considering small ($d = .2$), medium ($d = .5$), and large ($d = .8$) hypothesized effects in line with Cohen's (1988) conventions. Following Miller and Ulrich (2020), the Type-1 error probability was set to the standard level $\alpha = .05$, and the nominal power to detect the hypothesized effect was $1 - \beta = .90$ for either procedure. For the ISP, again following the recommendations of Miller and Ulrich (2020), the optimal maximum number of segments $k_{\max}$ and the optimal α_{strong} parameter was determined by numerical search to minimize the expected sample size given that the base rates for H_0 and H_1 are equal (that is, $p(H_0) = p(H_1) = .50$). For each of the $4 \times 3 = 12$ possible settings of the simulations, 10,000 Monte-Carlo replications of both sequential procedures were performed.

Results and Discussion

Table 1 summarizes the observed error rates of the two procedures depending on the simulation setting. As expected, both procedures keep their nominal $\alpha = .05$ level very well. Also, the observed Type-1 error rate of the SPRT is always less than .05, that is, the SPRT behaves somewhat conservatively. This conservative behavior occurs because the SPRT procedure almost always clearly exceeds the thresholds for accepting H_0 or H_1 rather than matching them exactly. This so-called "overshooting" is a well-known phenomenon in all types of SPRT applications. Consequently, the nominal α and β values serve as *upper bounds* to the actual error rates in the SPRT and rarely match them exactly (cf. Schnuerch & Erdfelder, 2020; Schnuerch et al., 2021).

Virtually the same picture emerges if we look at the observed error rates for H_1 scenarios with hypothesized population effect

[1] R scripts for all simulations and analyses as well as all simulated data are available at the Open Science Framework (https://osf.io/b4mhc/).

Table 1

Observed Error Rates of One-Tailed SPRT and ISP t Tests for Different True Population Effect Sizes δ (Rows) and Given Nominal Error Rates of α = .05 and β = .10, Separately for Different Hypothesized Population Effect Sizes d (Columns) in the Sequential Procedures

	Hypothesized effect size under H_1					
	d = 0.20		d = 0.50		d = 0.80	
True effect size	SPRT	ISP	SPRT	ISP	SPRT	ISP
δ = 0	.046	.051	.044	.049	.039	.050
δ = 0.20	**.092**	**.107**	.678	.694	.842	.833
δ = 0.50	.000	.000	**.082**	**.095**	.376	.421
δ = 0.80	.000	.000	.003	.003	**.079**	**.100**

Note. Error rates set in boldface type refer to simulations with correctly specified effect sizes under H_1.

sizes specified correctly (i.e., $d = \delta$, see the error rates set in boldface type in Table 1). In these cases, the nominal error rate $\beta = .10$ is never exceeded (apart from a single exception caused by sampling error). Just like in the H_0 scenario, overshooting leads to SPRT error rates somewhat smaller than $\beta = .10$. Hence, the actual power is slightly larger than the prespecified $1 - \beta = .90$. Not surprisingly, when H_1 holds but the hypothesized H_1 population effect size d is misspecified, actual Type-2 error rates are almost zero when d undercuts the true effect size ($d < \delta$) and much larger than .10 when d overbids it ($\delta < d$). This holds for both the SPRT and ISP, with negligible differences between procedures.

Most importantly in the present context, Table 2 summarizes the average sample sizes required for each procedure to reach a decision. As expected, the average N is always smaller for the SPRT compared with the ISP, with the single exception of an incorrectly assumed large effect under H_1 ($d = .80$) when the true effect is of medium size ($\delta = .50$). The gain in efficiency achieved by using the SPRT is often considerable, the more so the smaller δ and d. Under H_0, the relative efficiency gain in average sample sizes can be as high as 25% if the hypothesized population effect size is $d = .20$. Under H_1, up to 20% smaller average sample sizes are possible if $\delta = d = .50$. In sum, the SPRT can save up to 25% of the expected sample size required for the ISP. It is safe to say that the efficiency gain is larger than 10% on average, an advantage that is comparable to the relative efficiency gain of the SPRT compared with group-sequential designs with four predefined looks (Schnuerch & Erdfelder, 2020).

To illustrate, Figure 1 depicts the distribution of sample sizes required for the SPRT (dark grey) and the ISP (light grey) to reach a decision based on the $\delta = d = .50$ simulation setting with 10,000 Monte Carlo samples. In this case, $k_{\max} = 5$ is an optimal maximum number of segments for the ISP to minimize the expected sample size. Hence, the ISP histogram is discrete with five possible sample sizes until the procedure will eventually reach a decision.

Notably, although both the SPRT and the ISP are clearly more efficient on average than the corresponding fixed-N, one-tailed Neyman-Pearson t test, it can happen that the sequential procedures require more observations than the Neyman-Pearson test. For the settings underlying Figure 1, an a priori power analysis results in a required fixed-N sample size of $N_{\mathrm{NP}} = 140$ (Faul et al., 2009). As can be seen in Figure 1, this N_{NP} is exceeded by the SPRT in about 14% of the cases and by the ISP in about 16% of the cases. Results are not much different for other simulation settings (see Table 3). Thus, although we know the maximum possible sample size $N_{\max}$ beforehand in the ISP, the actual N of the ISP at termination will exceed N_{NP} for given α, β, and hypothesized effect sizes about as often as the corresponding N of the SPRT does.

General Discussion

The present study extends previous results of Schnuerch and Erdfelder (2020) by showing that the SPRT is not only more efficient than sequential Bayes factors and certain group-sequential designs for a wide range of scenarios but also more efficient than the ISP recently proposed by Miller and Ulrich (2020), especially when hypothesized population effect sizes are small. This result is not surprising because it has been well known since Wald's (1947) and Wald and Wolfowitz's (1948) seminal publications that the SPRT provides maximum efficiency for tests between two simple hypotheses.

Table 2

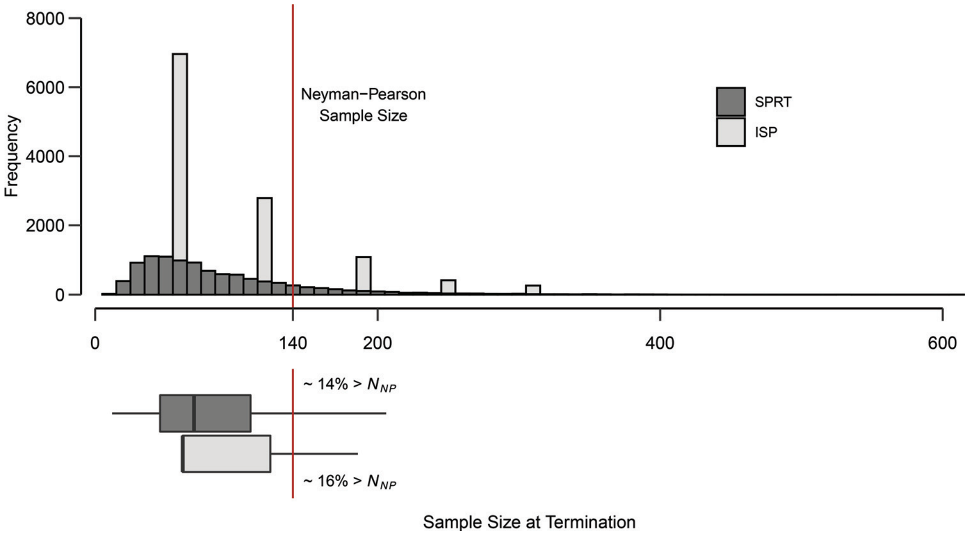
Average Sample Sizes of One-Tailed SPRT and ISP t Tests for Different True Population Effect Sizes δ (Rows) and Nominal Error Rates of α = .05 and β = .10, Separately for Different Hypothesized Population Effect Sizes d (Columns) in the Sequential Procedures

	Hypothesized effect size under H_1					
	d = 0.20		d = 0.50		d = 0.80	
True effect size	SPRT	ISP	SPRT	ISP	SPRT	ISP
δ = 0	414 (316, 536, 1060)	549 (382, 764, 1146)	73 (56, 94, 184)	89 (62, 124, 186)	31 (24, 38, 76)	36 (26, 52, 78)
δ = 0.20	**500 (408, 642, 1186)**	**628 (382, 764, 1528)**	118 (88, 154, 312)	121 (62, 186, 310)	43 (32, 56, 112)	44 (26, 52, 104)
δ = 0.50	153 (142, 182, 264)	383 (382, 382, 382)	**87 (70, 110, 200)**	**103 (62, 124, 248)**	52 (40, 68, 134)	51 (52, 78, 130)
δ = 0.80	92 (88, 106, 138)	382 (382, 382, 382)	49 (42, 58, 92)	69 (62, 62, 124)	**39 (32, 48, 88)**	**42 (26, 52, 78)**

Note. The values depicted in parentheses indicate 50%, 75%, and 95% percentiles, respectively, of the total sample size distribution. Boldface numbers indicate simulations with correctly specified effect sizes under H_1.

Figure 1

Distribution of Total Sample Sizes for the One-Tailed SPRT and ISP Sequential t Tests Based on Nominal Error Rates $\alpha = .05$ and $\beta = .10$ for a Hypothesized Medium Effect Size Under H_1 That Matches the True Effect Size in the Underlying Population (i.e., $d = \delta = .50$)



Note. The upper part of the figure shows Monte-Carlo-estimated histograms of sample sizes at termination (with the Neyman-Pearson sample size N_{NP} = 140 marked on the abscissa) and the lower part depicts corresponding box plots. See the online article for the color version of this figure.

What are the implications of this result for the choice between different sequential test strategies? If the costs of statistical hypothesis testing are proportional to the number of observations required to reach a statistical decision, SPRTs should be used whenever an a priori known maximum sample size is not mandatory. We believe this is quite often the case in behavioral research, for example, in direct replications of theoretically important original studies (cf. Erdfelder & Ulrich, 2018), whenever online experiments or surveys are conducted without a definite time limit, or when hypotheses are tested for single participants by repeatedly sampling test items or test stimuli from very large populations (e.g., Schnuerch et al., 2020). We agree with Lakens (2014, p. 701) that researchers have an ethical obligation to engage in most efficient sequential procedures to "avoid waste of the time of participants and the money provided by taxpayers." The SPRT is ideally suited to serve this goal whenever an a priori defined deadline for finishing data collection is not required. Also, in contrast to Miller and Ulrich's (2020, p. 3) claim, practical aspects of research planning are easily reconciled with SPRT applications. Both lab resources and the budget required to pay participants can be planned a priori, simply by calculating the expected distribution of sample sizes (see Figure 1) and by multiplying the expected sample size by a factor that represents the costs per observation (or per participant), plus a security margin to cope with effects of sampling error. Overall, therefore, researchers will save lab and budget resources by employing the SPRT rather than any other sequential or nonsequential test procedure that controls for α and β.

Concerning simplicity of computations, employing the SPRT is not a big problem either. The two-groups SPRT *t* test (Hajnal, 1961) essentially involves repeated calculations of likelihood ratios of observed *t* statistics under the noncentral and the central *t* distribution, respectively. Given the availability of

Table 3

Relative Frequency of SPRT and ISP Simulations Exceeding the Corresponding Neyman-Pearson Sample Size N_{NP} Determined by an a Priori Power Analysis, Separately for Different True (δ) and Hypothesized Population Effect Sizes (d)

	Hypothesized effect size under H_1					
	d = 0.20		*d* = 0.50		*d* = 0.80	
True effect size	SPRT	ISP	SPRT	ISP	SPRT	ISP
δ = 0	.095	.094	.108	.094	.112	.080
δ = 0.20	**.130**	**.158**	.288	.250	.244	.163
δ = 0.50	.000	.000	**.143**	**.161**	.326	.249
δ = 0.80	.000	.000	.005	.009	**.170**	**.138**

Note. Numbers set in boldface type refer to simulations with correctly specified effect sizes under H_1 ($d = \delta$).

noncentral t distribution functions in free software packages such as R, implementation of the SPRT t test involves just a few lines of R code (as detailed in Schnuerch & Erdfelder, 2020, p. 210; a workable, user-friendly R script may be downloaded from https://osf.io/wz8da/). As an alternative, we also provide a shiny app that handles various SPRT t tests for different H_1 models, not just the case of a simple point H_1 as discussed here (https://martinschnuerch.shinyapps.io/Waldian-t-Tests/). Note that, if desired, the SPRT can also be conducted with a desk calculator, for example, by repeatedly calculating probability density ratios for observed t values based on the nctpdf(t, df, nc) and the tpdf(t, df) functions, respectively, implemented in the G*Power calculator (Faul, 2020, pp. 7–8; Faul et al., 2009).

One should keep in mind, however, that the general theory of the SPRT only applies to tests between simple hypotheses, that is, when the likelihood functions are fully specified. Given that most of the tests conducted in psychology refer to composite hypotheses, this is a potential limitation. Fortunately, techniques have become available to cope with this limitation in many situations, for example, replacing composite hypotheses by simple hypotheses on transformations of the data (Cox, 1952; Rushton, 1950; Schnuerch & Erdfelder, 2020), using weight functions and integrating out nuisance parameters (Wald, 1947; see also Schnuerch et al., 2021), or replacing unknown nuisance parameters by their maximum likelihood estimates for a given set of data—an asymptotic method that works nicely when initial sample sizes in the SPRT are not too small (Cox, 1963; Schnuerch et al., 2020).

However, if it is mandatory to know the maximum possible sample size in advance, the ISP is a viable alternative to the SPRT. The ISP compares very well in terms of mathematical foundation, simplicity, applicability, precision, and efficiency with established group-sequential methods. In addition, its deficits in efficiency compared with the SPRT are relatively small when to-be-detected effects under H_1 are large (see Table 2). Moreover, the ISP has several other advantages compared with previously proposed sequential methods that have been nicely summarized by Miller and Ulrich (2020).

There is one aspect, however, in which we respectfully disagree with Miller and Ulrich (2020). They point out as an advantage of the ISP that it is "usable without assuming an effect size d" (Table 1 in Miller & Ulrich, 2020). Their claim is correct since it is possible to control for α in the ISP without any hypothesis about δ—much like in null hypothesis significance testing (NHST). However, we question the premise that this constitutes an advantage. By blindly proceeding along these lines without considering population effect size and power, the ISP would inherit the well-known major disadvantages of NHST such as running the risk of performing severely underpowered studies and, in turn, wasting efficiency. Defining the procedure by choosing an arbitrary sample size n_s essentially is an implicit commitment to an arbitrary population effect size—the unknown effect for which ISP has sufficient statistical power. Only the explicit specification of a population effect size enables researchers to exploit the full strength of the ISP: The joint optimization of all input parameters such that the expected sample size for detecting the hypothesized effect with the desired power becomes a minimum (cf. Figure 5 in Miller & Ulrich, 2020). Reasonable candidates for a priori effect size specifications are (a) the minimum population effect of interest in a given context (see, e.g., Lakens, 2014) or (b) a meta-analytic effect size estimate based on previous related studies that takes publication bias into account (cf. Ulrich et al., 2018).

Although the expected sample size in the ISP can be notably larger than in the SPRT, its maximum number of observations is known in advance and, as already shown by Miller and Ulrich (2020), the ISP is much more efficient on average than the corresponding fixed-N Neyman-Pearson test. Obviously, there is a price to pay for this efficiency gain relative to the Neyman-Pearson test, and this price is that $N_{\max}$ of the ISP will exceed N_{NP}. Thus, the importance of knowing $N_{\max}$ before starting the sequential procedure should not be overemphasized. Samples considerably larger than N_{NP} can occur with both the SPRT and the ISP for given α, β, and hypothesized effect size, but fortunately such cases occur rarely in practice.

In sum, both the SPRT and the ISP are useful tools in the sequential statistics toolbox, with the SPRT serving as the default choice whenever it is not necessary to specify $N_{\max}$ in advance. The ISP, however, is a reasonable alternative whenever $N_{\max}$ needs to be known in advance.

References

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum. https://doi.org/10.1016/C2013-0-10517-X

Cox, D. R. (1952). Sequential tests for composite hypotheses. *Mathematical Proceedings of the Cambridge Philosophical Society*, *48*(2), 209–226. https://doi.org/10.1017/S030500410002764X

Cox, D. R. (1963). Large sample sequential tests for composite hypotheses. *Sankhya: The Indian Journal of Statistics, Series A (1961–2002)*, *25*(1), 5–12.

Erdfelder, E., & Ulrich, R. (2018). Zur Methodologie von Replikationsstudien [On the methodology of replication studies]. *Psychologische Rundschau*, *69*(1), 3–21. https://doi.org/10.1026/0033-3042/a000387

Faul, F. (2020). *G*Power 3.1 manual*. https://www.psychologie.hhu.de/fileadmin/redaktion/Fakultaeten/Mathematisch-Naturwissenschaftliche_Fakultaet/Psychologie/AAP/gpower/GPowerManual.pdf

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analysis using GPower 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Hajnal, J. (1961). A two-sample sequential t test. *Biometrika*, *48*(1–2), 65–75. https://doi.org/10.1093/biomet/48.1-2.65

Lai, T. L. (1981). Asymptotic optimality of invariant sequential probability ratio tests. *Annals of Statistics*, *9*(2), 318–333. https://doi.org/10.1214/aos/1176345398

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

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

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, *64*(2), 191–199. https://doi.org/10.1093/biomet/64.2.191

Proschan, M. A., Lan, K. K. G., & Wittes, J. T. (2006). *Statistical monitoring of clinical trials: A unified approach*. Springer. https://doi.org/10.1007/978-0-387-44970-8

Rushton, S. (1950). On a sequential t test. *Biometrika*, *37*(34), 326–333. https://doi.org/10.2307/2332385

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

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

Schnuerch, M., Heck, D. W., & Erdfelder, E. (2021). *Waldian t tests: Sequential Bayesian t tests with controlled error probabilities* [Manuscript submitted for publication]. https://doi.org/10.31234/osf.io/x4ybm

Ulrich, R., Miller, J., & Erdfelder, E. (2018). Effect size estimation from t statistics in the presence of publication bias: A brief review of existing approaches with some extensions. *Zeitschrift für Psychologie*, *226*(1), 56–80. https://doi.org/10.1027/2151-2604/a000319

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

Wald, A., & Wolfowitz, J. (1948). Optimum character of the sequential probability ratio test. *Annals of Mathematical Statistics*, *19*(3), 326–339. https://doi.org/10.1214/aoms/1177730197

Zhu, L., Ni, L., & Yao, B. (2011). Group sequential methods and software applications. *The American Statistician*, *65*(2), 127–135. https://doi.org/10.1198/tast.2011.10213

Received November 23, 2020
Revision received March 8, 2021
Accepted March 10, 2021 ■