European Journal of Social Psychology, *Eur. J. Soc. Psychol.* **44**, 701–710 (2014)
Published online in Wiley Online Library (wileyonlinelibrary.com) **DOI**: 10.1002/ejsp.2023

Special issue article: Methods and statistics in social psychology: Refinements and new developments

Performing high-powered studies efficiently with sequential analyses

DANIËL LAKENS*
Human Technology Interaction Group, Eindhoven University of Technology, Eindhoven, The Netherlands

Abstract

Running studies with high statistical power, while effect size estimates in psychology are often inaccurate, leads to a practical challenge when designing an experiment. This challenge can be addressed by performing sequential analyses while the data collection is still in progress. At an interim analysis, data collection can be stopped whenever the results are convincing enough to conclude that an effect is present, more data can be collected, or the study can be terminated whenever it is extremely unlikely that the predicted effect will be observed if data collection would be continued. Such interim analyses can be performed while controlling the Type 1 error rate. Sequential analyses can greatly improve the efficiency with which data are collected. Additional flexibility is provided by adaptive designs where sample sizes are increased on the basis of the observed effect size. The need for pre-registration, ways to prevent experimenter bias, and a comparison between Bayesian approaches and null-hypothesis significance testing (NHST) are discussed. Sequential analyses, which are widely used in large-scale medical trials, provide an efficient way to perform high-powered informative experiments. I hope this introduction will provide a practical primer that allows researchers to incorporate sequential analyses in their research. Copyright © 2014 John Wiley & Sons, Ltd.

Repeatedly analyzing results while data collection is in progress has many advantages. Researchers can stop the data collection when observed differences reach a desired confidence level or when unexpected data patterns occur that warrant a reconsideration of the aims of the study. When, after an interim analysis, the effect is smaller than expected, researchers might decide to collect more data or even stop collecting data for specific conditions. One could easily argue that psychological researchers have an ethical obligation to repeatedly analyze accumulating data, given that continuing data collection whenever the desired level of confidence is reached, or whenever it is sufficiently clear that the expected effects are not present, is a waste of the time of participants and the money provided by taxpayers. In addition to this ethical argument, designing studies that make use of sequential analyses are more efficient compared with not performing sequential analyses. Incorporating sequential analyses into the study design can easily reduce the sample size of studies by 30% or more.

In psychology, sequential analyses are rarely, if ever, used. In recent years, researchers have been reminded of the fact that repeatedly analyzing data, and continuing the data collection when results are not significant, increases the likelihood of a Type 1 error, or a significant test result in the absence of any differences in the population (e.g., Simmons, Nelson, & Simonsohn, 2011). Flexibility in data collection does not lead to an inevitable increase in Type 1 errors. It is possible to repeatedly analyze data, and decide whether to continue or end the data collection on the basis of the significance levels of the accumulated data, while controlling the Type 1 error rate. Such interim analyses are common practice in large medical trials, where the continued data collection can be a matter of life and death when a treatment has unexpected negative consequences. The basic idea of sequential analyses has been developed in the 20th century (e.g., Armitage, McPherson, & Rowe, 1969; Dodge & Romig, 1929), and advances in these techniques over the last decennia have provided statistical procedures for sequential analyses that allow great flexibility while carefully controlling Type 1 error rates. Accessible mathematical introductions to sequential analyses for clinical trials can be found in books by Proschan, Lan, and Wittes (2006); Chow, Shao, and Wang (2003); and Jennison and Turnbull (2000), among others.

I believe sequential analyses are relevant for psychological science. There is an increasing awareness that underpowered studies in combination with publication bias (the tendency to only accept manuscripts for publication that reveal statistically significant findings) yield a scientific literature that potentially consists of a large number of Type 1 errors (e.g., Button et al., 2013; Ioannidis, 2005; Lakens & Evers, in press). There are several ways to increase the statistical power of a study (or the probability that a significant effect will be observed in a sample if the effect truly exists in the population), but the way that is easiest to control by researchers is to increase the sample size of their studies.

*Correspondence to: Daniël Lakens, Human Technology Interaction Group, Eindhoven University of Technology, IPO 1.33, PO Box 513, 5600 MB Eindhoven, The Netherlands.
E-mail: D.Lakens@tue.nl

Received 30 September 2013, Accepted 11 March 2014

Researchers have started to realize that especially in between-subject designs, a much larger number of participants have to be collected than psychologists were accustomed to if they desire to perform high-powered studies. Some researchers have suggested a minimum of 20 participants per condition (e.g., Simmons et al., 2011, later increased to a minimum of 50 participants per condition, Simmons, Nelson, & Simonsohn, 2013), but such well-intended suggestions are bad advice. Sample sizes should be determined on the basis of the expected size of the effect, the desired power, and the planned alpha level of the statistical test or on the basis of the desired width of the confidence interval (CI) surrounding the effect size estimate. An important question is how we are going to be able to increase sample sizes without greatly reducing the number of experiments one can perform. Sequential analyses provide a partial solution to this problem.

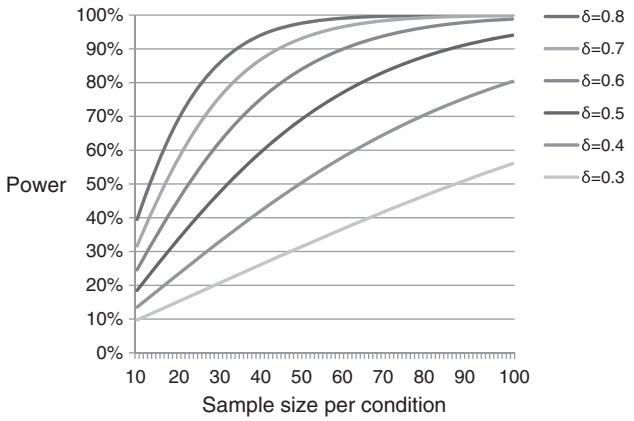

Figure 1. Statistical power as a function of the effect size (Cohen's δ) and sample size per condition

PRACTICAL ISSUES WHEN DESIGNING AN ADEQUATELY POWERED STUDY

One problem with planning the sample size on the basis of the size of an effect (as is done in an *a priori* power analysis) is that the effect size is precisely the information that the researcher is trying to uncover by performing the experiment. As a consequence, there is always some uncertainty regarding the required sample size needed to observe a statistically significant effect. Nevertheless, *a priori* power analyses are often recommended when designing studies to provide at least some indication of the required sample size (e.g., Lakens, 2013), and researchers therefore need to estimate the expected effect size when designing a study. One approach to obtain an effect size estimate is to perform a pilot study. To provide a reasonably accurate effect size estimate, a pilot study must already be quite large (e.g., Lakens & Evers, in press), somewhat surpassing their usefulness. A second approach is to base the effect size estimate on an effect size observed in a highly related study, while acknowledging that effect sizes might vary considerably because of the differences between the studies. Regardless of how effect sizes are estimated, estimated effect sizes have their own CIs (as any other sample statistic) and should be expected to vary between the lower and upper confidence limits across studies.

Because statistical power is a function that increases concave down (especially for larger effect sizes, see Figure 1), improving the power of a study from 0.8 to 0.9 requires a larger increase in same size than is needed to improve the power of a study from 0.4 to 0.5. Reducing the Type 2 error (not observing a statistically significant effect in a sample, when the effect exists in the population) becomes increasingly costly the higher the desired power and the smaller the expected effect size. A widely used recommendation is to aim for a minimum power of 0.8 (Cohen, 1988). This still leaves a one in five chance of not finding an effect that actually exists, so a higher power (e.g., 0.9 or 0.95) is often desirable. Whenever the sample size recommended by an *a priori* power analysis has been collected, and the performed statistical test reveals a non-significant result (due to an underestimation of the effect size, random variation of the effect size, or both), the researcher is faced with a dilemma. The researcher could replicate the study (which requires substantial resources) and report a small-scale meta-analysis (e.g., Cumming, 2012). Alternatively, the researcher might be tempted to continue the data collection, which increases the Type 1 error rate (Simmons et al., 2011). Sequential analyses give a researcher more flexibility to continue the data collection and stop when an interim analysis reveals a significant difference.

Because the power function increases concave downwards (Figure 1), there is a reasonably high chance (often >50%) to observe a significant effect after collecting only half the number of participants suggested by an *a priori* power analysis. Using sequential analyses thus often allows researchers to terminate the data collection before the sample size recommended by an *a priori* power analysis is reached, and is therefore more efficient when designing studies with the goal to demonstrate a statistically significant effect.

The idea that we need to collect large amounts of data without any flexibility worries researchers, and some researchers have argued against a fixation on Type 1 error control. Ellemers (2013, p. 3) argues "we are at risk of becoming methodological fetishists," which "would reverse the means and the ends of doing research and stifles the creativity that is essential to the advancement of science." Although flexibility in the generation of hypotheses is, in principle, completely orthogonal to how strict these hypotheses are tested empirically, there is a real risk that researchers will become more conservative in the ideas they test. If researchers believe they should perform high-powered experiments with large samples without looking at the data until all participants have been collected, they might not pursue hypotheses that initially seem more unlikely.

Murayama, Pekrun, and Fiedler (2013) discuss the practice of adding additional observations to a study, on the basis of the observed *p*-value, and warn against jumping to the extreme conclusion that continuing data collection after analyzing the data should be banned. They examine what happens when researchers collect additional observations *only* when an analysis reveals a *p*-value between .05 and .10. Such a practice, they show, would lead to a modest increase in Type 1 error rates (as long as the number of times additional data are collected is limited). Although this is important to realize, underpowered studies will often yield *p*-values higher than .10 when there is a real effect in the population. Because using

sequential analyses is not very complex, it is preferable to know and use procedures that control Type 1 error rate while performing interim analyses. In the remainder of this article, I will explain how Type 1 error control is possible and provide a practical primer on how to perform sequential analyses.

TYPE 1 ERROR CONTROL WHILE PERFORMING INTERIM ANALYSES

Statistical procedures to perform sequential interim analyses while data collection is still in progress have been available for a long time (e.g., Armitage et al., 1969; Dodge & Romig, 1929). The main idea is straightforward. In a study without interim analyses, a statistical test is performed when all the data have been collected. With a symmetrical two-sided test, and an $\alpha = .05$, this test should yield a *Z*-value larger than 1.96 (or smaller than −1.96) for the observed effect to be considered significant (which has a probability smaller than .025 for each tail, assuming the null hypothesis is true). When using sequential analyses with a single-planned interim analysis, and a final analysis when all data are collected, one test is performed after *n* (e.g., 80) of the planned *N* (e.g., 160) observations have been collected, and another test is performed after all *N* observations are collected. This means a researcher plans to reject the null hypothesis either when after *n* observations Z_n is sufficiently high (or low) or when after *N* observations Z_N is sufficiently high (or low). As Leifer and Geller (2012) explain, this means that to control the probability of a Type 1 error under the null hypothesis, we need to select boundary critical *Z*-values c_1 and c_2 (for the first and the second analysis) such that (for the upper boundary) the probability (*Pr*) that the null hypothesis is rejected either when in the first analysis $Z_n \geq c_1$ or (when $Z_n < c_1$ in the first analysis) when $Z_N \geq c_2$ in the second analysis. In formal terms,

$$Pr\{Z_n \geq c_1\} + Pr\{Z_n < c_1, Z_N \geq c_2\} = 0.025 \qquad (1)$$

One-sided tests or asymmetric boundaries are sometimes used (see Proschan et al., 2006), and with more than one interim analysis, additional critical values have to be determined following the same rationale. There are different ways in which these critical values can be established. The Pocock boundary (Pocock, 1977) increases the critical value (or lowers the alpha level) to the same value for each interim analysis such that the overall alpha level remains .05. For example, for two-planned analyses (one interim and one after all data are collected), the *p*-value would become .0294 for each analysis. The O'Brien-Fleming procedure differs from the Pocock boundary in that it sets a high critical value for the first interim analysis, when the variability in the data is relatively large, but sets a critical value for the final analysis that is closer to a study without interim looks. One limitation of the Pocock and O'Brien-Fleming boundaries is that they require that the number of interim analyses is planned in advance and that these analyses are performed with an equal number of observations between each look.

This lack of flexibility is impractical in medical settings, where data and safety monitoring boards meet at fixed times each year, and it is not always feasible to control the number of patients between these meetings. In psychological research, it might be difficult to pause an experiment after a predefined number of observations have been collected and wait for the data analysis to be performed. Lan and DeMets (1983) developed the idea of a *spending function* that specifies how much alpha to use at which time in the trial. Spending functions do not require the number of interim analyses to be specified in advance (as long as the *Z*-value at one interim analysis does not determine the time until the following analysis). Pocock-like and O'Brien-Fleming-like alpha spending functions, which are continuous approximations of the Pocock and O'Brien-Fleming boundaries, can be calculated using a freeware program called WinDL created by Reboussin, DeMets, Kim, and Lan (2000), the GroupSeq package in R, or commercial software packages such as PASS. In addition to the Pocock and O'Brien-Fleming functions, a linear spending function (a power family function, see Jennison & Turnbull, 2000) is often used, which lies between the Pocock and O'Brien-Fleming boundaries and spends the alpha level continuously over the interim analyses.

Stopping a trial early has clear benefits, such as saving time and money when the available data are considered convincing, but it also has disadvantages, as noted by Pocock (1992). Effect size estimates from small studies are sometimes still "sailing the seas of chaos" (Lakens & Evers, in press), and the large variation in effect size estimates makes results from small studies less convincing, because the study might have stopped at a "random high." Obviously, this argument holds for any small study, regardless of whether or not sequential analyses were used, and because sequential analyses are performed with a lower alpha level, and effect size estimates are adjusted for bias when the data collection is terminated early, sequential analyses provide more reliable effect size estimates than traditional small-scale studies. Another disadvantage is that small samples typically yield effect size estimates with extremely wide CIs, which is generally undesirable. These issues are inherent limitations of null-hypothesis significance testing (see Discussion). It is therefore recommended to use sequential analyses and significance testing to identify which effects in a line of research hold promise. Follow-up studies with larger sample sizes, or meta-analyses, can be used to provide more accurate effect size estimates.

It is also possible that the effect size estimate at an interim analysis is (close to) zero, which indicates that the effect is non-existent or very small. Because the chance of observing a statistically significant difference is very small or would require huge sample sizes, researchers can decide to terminate the data collection early for *futility* to spare time and resources. Obviously, researchers might want to continue a study even when the conditional or predictive power after an interim analysis is very low; for example, when they are also interested in demonstrating that there are no effects in a study. In many situations, the decision to stop a study for futility will be more complex, and I will return to this issue when discussing how to define a smallest effect size of interest.

An Illustrative Example When a Reliable Effect Size Estimate Exists

As an example of how sequential analyses can be used, suppose a researcher is interested in whether a relatively

well-established effect in The Netherlands can also be observed in Japan. A meta-analysis of the studies performed in The Netherlands has yielded a meta-analytic effect size estimate of Cohen's $d = 0.5$, and he or she expects that the effect should generalize to the Japanese population. Because of practical constraints, the researcher determines that he or she is willing to collect a maximum of 180 observations, which means the study has a power of 0.92 to observe an effect of $d_{pop} = 0.5$ in a single statistical test performed after all participants are collected. Although he or she is willing to collect up to 180 observations, he or she would prefer to stop earlier if the available data provide clear support for his or her hypothesis, and he or she would also prefer to stop earlier if it seems unlikely that the effect is larger than $d = 0.4$.

The researcher decides to perform two-sided interim analyses after collecting 60, 120, and 180 participants. Using a linear spending function (in WinDL indicated by the power family function with a phi of 1, in PASS indicated by the (alpha) (time) spending function, and in GroupSeq indicated by the alpha*t^phi function), we can calculate alpha boundaries for the three analyses (two interim and one final) of .017, .022, and .028, or Z boundaries of 2.394, 2.294, and 2.200. A detailed step-by-step guide to calculating the boundaries using the GroupSeq package in R or the WinDL software by Reboussin et al. (2000) is provided in the supplementary materials available at http://osf.io/uygrs/files/ Let us assume that after collecting 30 participants in each condition, the first interim analysis reveals a Cohen's $d_s = 0.51$, 95% CI [−0.004, 1.025], $t(58) = 1.985$, $p = .052$.[1] The effect does not fall below the boundary value of 0.017, and the researcher continues the data collection. After 60 participants in each condition, the second interim analysis reveals a Cohen's $d_s = 0.46$, 95% CI [0.09, 0.82], $t(118) = 2.49$, $p = .014$. This falls below the significance level of 0.022, and the data collection is terminated.

The use of sequential analyses reduces the statistical power of the study to 0.89. In 32.37% of the studies, the researcher would find an effect after collecting 60 participants; in 36.70% of the studies, he or she would find an effect after 120 participants; and in 19.94% of the studies, he or she would find an effect after 180 participants (for a total power of 0.89). In 10.99% of the studies, on average, he or she would not find an effect, even though it exists. In 100 studies, this researcher would need to collect an average of 11 940 participants (60 participants in 32 successful studies, 120 participants in 37 successful studies, 180 participants in 20 successful studies, and 180 participants in 11 studies that yield a Type 2 error). Let us compare this with non-sequential procedures. The study without interim analyses will not suffer the reduction in power due to sequential testing, so that the same level of power (0.89) can be achieved by collecting 166 participants in each of the 100 studies, for a total of (100×166) 16 600 participants. Despite the slightly lower sample size to achieve the same level of statistical power in a study that does not make use of sequential analyses compared with a study that uses sequential analyses, the use of sequential analyses allows researchers to perform well-powered experiments while collecting approximately 28% fewer participants because of the possibility to terminate the data collection earlier.

One might have thought that given the availability of a reasonably accurate effect size estimate, and the reduction in power by incorporating interim analyses in the study design, the researcher should just perform a well-powered study without interim analyses. However, because power increases in a concave downwards function (Figure 1), there is a reasonably high chance that a researcher can terminate the data collection before the sample size recommended by an *a priori* power analysis is reached. Although a power of 0.50 would not be satisfactory for the final analysis that is performed in a study, having a power of 0.5 at an interim analysis means there is a 50% chance to observe an expected effect and terminate the data collection. Using sequential analyses, a researcher benefits from the possibility to terminate the data collection early, while being able to continue the data collection when the data are not yet statistically significant.

As an example, Table 1 displays the average sample sizes for studies examining true effects using four sequential analyses (which takes into account the reduction in the required sample size due to early stopping) using a linear spending function, compared with the required sample size for a study that does not use sequential analyses, both designed to achieve a power of either 0.80 or 0.90 (alpha = .05), for a range of five effect sizes (Cohen's $d_{pop} = 0.2$, 0.3, 0.43, 0.5, and 0.8). The average reduction in the number of participants required to yield a statistically significant result using sequential analyses compared with non-sequential analyses is around 20% for studies designed to achieve a power of 0.8, and around 30% higher for studies designed to achieve a power of 0.9. The exact efficiency benefit will differ depending on the number and timing of the interim analyses, and the goal to design an efficient study should be less important than designing a study with a high informational value. For example, researchers might not want to perform an interim analysis after 25% of the planned sample is collected, when the sample size will often be too low to yield reliable inferences. Nevertheless, sequential analyses will be more efficient in practically all situations.

An Illustrative Example When No Reliable Effect Size Estimate Exists

When examining a completely novel hypothesis, a researcher might not have a precise idea of the expected effect size. In such situations, sequential analyses are especially useful. Let us imagine a researcher who wants to examine the novel idea that people who speak in a louder voice do so to compensate for feelings of insecurity. He or she manipulates feelings of insecurity by asking participants in the experimental group to think back to a time they felt very insecure about themselves (while participants in the control group think back to a time when they were watching television) and then asks participants to read a piece of text out loud. He or she expects that the manipulation will make participants in the experimental group feel less secure about themselves and speak in a louder voice when reading the text out loud. Without any idea of the

[1]We can also consider a scenario where the interim analysis returns a p-value that falls between the traditional significance level (0.05) and the boundary value for sequential analyses (e.g., 0.017). After observing a p-value of 0.048, a researcher might regret having decided to use sequential analyses. Nevertheless, sequential analyses are on average more efficient, p-values above .03 are never very strong support for a hypothesis (Lakens & Evers, in press), and the researcher can perhaps find solace in the knowledge that collecting more data always increases the informational value of studies.

Table 1. Average sample size for sequential (four analyses, linear spending function) and non-sequential studies designed to achieve a power of 0.80 or 0.90 (alpha = .05) for five effect sizes, and the reduction in the average number of collected observations for an independent *t*-test

	80% power			90% power		
	Sequential *N*	Non-sequential *N*	Reduction (%)	Sequential *N*	Non-sequential *N*	Reduction(%)
$\delta = 0.8$	39.17	52	24.67	45.91	68	32.49
$\delta = 0.5$	100.82	128	21.23	117.88	172	31.47
$\delta = 0.43$	135.73	172	21.09	158.82	230	30.95
$\delta = 0.3$	278.38	352	20.91	326.33	470	30.57
$\delta = 0.2$	625.09	788	20.67	733.30	1054	30.43

effect size to expect in this novel line of research, he or she decides that he or she is willing to collect at most 400 participants, if the effect reliably exists. However, he or she would prefer to stop collecting data when the effect is reliably observed, or when it seems reasonably certain that the effect does not exist. He or she decides to terminate the data collection if the effect size estimate is lower than 0.30 during an interim analysis, for which he or she has a 0.85 power in a non-sequential analysis given the sample size of 400 participants and a 0.80 power for three sequential analyses using a Pocock spending function.

Because there is no way to know the true effect size, let us imagine what might happen if the effect is small ($d = 0.3$), medium ($d = 0.5$), or large ($d = 0.8$) when the researcher performs interim analyses after $n = 100$, $n = 200$, and $N = 400$ participants (remember that the times at which the analyses are performed do not have to be evenly spaced when using an alpha spending function, as in this example where analyses are performed at time = 0.25, 0.50, and 1). Table 2 summarizes the expected power of the test (and the nominal alpha level for the test using the Pocock spending function) depending on the size of the effect. The Pocock spending function is used because it distributes the alpha level evenly over all interim analyses, which is a justifiable choice given the uncertainty about the expected effect size. Acceptable levels of power (>0.80) are reached at the first look for large effects, at the second look for medium effects, and at the final look for small effects. Let us assume that after collecting 50 participants in each condition, the first interim analysis reveals a Cohen's $d_s = 0.55$, 95% CI [0.15, 0.95], $t(98) = 2.76$, $p = .007$. Even though the effect size estimate is inaccurate (as indicated by the large CI), the *p*-value is lower than the alpha boundary of 0.018 (Table 2), which indicates that the data provide support for the hypothesis and the data collection can be terminated.

Table 2. Alpha level and power as a function of sample size and effect size (δ) for sequential and non-sequential tests

Sample size	Nominal alpha	Power ($\delta = 0.3$)	Power ($\delta = 0.5$)	Power ($\delta = 0.8$)
		Sequential tests		
100	0.018	0.193	0.552	0.949
200	0.018	0.435	0.889	1.000
400	0.026	0.800	0.998	1.000
		Non-sequential tests		
100	0.050	0.318	0.697	0.977
200	0.050	0.560	0.940	1.000
400	0.050	0.849	0.999	1.000

Alternatively, let us assume that the first interim analysis reveals a Cohen's $d_s = 0.14$, 95% CI [−0.25, 0.54], $t(98) = 0.71$, $p = .477$. It is still possible that the true effect size is larger than $d = 0.30$, but on the basis of the current data, it is more likely to be lower. The researcher can therefore choose to terminate the data collection (and accept the risk that he or she is making a Type 2 error, given the uncertainty around the effect size estimate due to the relatively low sample size). The researcher may also decide to postpone the decision to terminate the experiment to the second interim analysis, because more data always yield more accurate inferences. Whereas the benefit of sequential analyses to design more efficient studies when the effect is true was highlighted earlier (Table 1), sequential analyses are also more efficient when examining hypotheses that are not true or when examining effects that are too small to be of interest to a researcher. Researchers should be aware that accurate effect size estimates require large samples, especially for small effect sizes, and early stopping often means it remains relatively probable that a small effect exists. Researchers should therefore not be tempted to conclude that there is no effect when terminating the data collection at an early interim analysis and instead interpret their non-significant data as indicating that the effect size is likely to be smaller than the effect size of interest to the researcher.

ADJUSTMENTS WHEN REPORTING TEST STATISTICS

Sequential analyses allow us to conclude whether a statistically significant effect is present or not, but in addition to this conclusion, researchers often want to report test statistics, such as a *p*-value, an effect size, and CIs around the difference. In non-sequential analyses, these test statistics all follow from the same theory, but this is not true for sequential analyses. A *p*-value is the probability of observing a result at least as extreme as the result that was observed, given that the null hypothesis is true. When using sequential analyses, more than one analysis is performed, and the definition of a result "at least as extreme" needs to be redefined. The recommended procedure to determine what "at least as extreme" means is to order the outcomes of a series of sequential analyses in terms of the stage at which the study was stopped, where earlier stopping is more extreme than later stopping, and where studies with higher *Z*-values are more extreme, when different studies are stopped at the same time (see Proschan et al., 2006). This is referred to as *stagewise ordering*. Finally, the

effect size estimate needs to be corrected for bias when stopping early, because the observed effect size at the moment the study is stopped could be an overestimation of the true effect size. Although procedures to control for bias have been developed, there is still much discussion about the interpretation of such effect sizes, and studies using non-adaptive designs, followed by a meta-analysis, might be needed if an accurate effect size estimate is paramount.

Let us continue the earlier example where an effect is examined in a Japanese population. After collecting 60 participants, the researcher performs an independent *t*-test and finds a non-significant difference in the expected direction, $t(58) = 1.26$, $p = .21$. At the second interim analyses after 120 participants, the difference is statistically significant, $t(118) = 2.65$, $p = .009$. Because the *p*-value at the second analysis is lower than the alpha boundary of .022, the researcher terminates the data collection. To calculate the *p*-value corrected for sequential analyses, we have to report the probability of observing a specific alpha level (as in a normal statistical test), *while not observing a significant difference at earlier interim analyses*. In this case, we report the probability that the *Z* boundary was not crossed at interim analysis 1 (when the boundary was $Z = 2.394$, see the first illustrative example) and that we observed the *Z*-value associated with the observed *p*-value in the second interim analysis.[2] We can again use WinDL or the GroupSeq package in R to calculate a *p*-value, which yields an adjusted $p = .023$. Step-by-step guides for WinDL and R (including screenshots and a spreadsheet that can be used to perform additional calculations) are available as supplementary material from http://osf.io/uygrs/files/ and can also be used to calculate adjusted mean differences, Cohen's *d*, and 95% CI for the observed difference.

INTERNAL PILOT STUDIES, CONDITIONAL POWER, AND ADAPTIVE DESIGNS

During an interim analysis, the observed results can be used to adapt the study design, for example, by increasing the total sample size that will be collected. According to the European Medicines Agency (2006), "a study design is 'adaptive' if statistical methodology allows the modification of a design element (e.g., sample-size, randomization ratio, number of treatment arms) at an interim analysis with full control of Type I error rate." With interim analyses, it becomes possible to use data collected early in a study as an *internal pilot study* (Wittes & Brittain, 1990) and use the effect size estimate from an interim analysis to determine the sample size for the full study. Small adaptations of the planned sample size might be based on differences in *nuisance parameters*, which are not the main parameters of interest in a study. For example, an interim analysis might reveal a difference between the expected standard deviation of a dependent variable and the observed standard deviation in the sample. If the observed standard deviation in the sample is larger than expected, more participants will be needed to maintain the desired statistical power of the test.

[2]This can be done in Excel using: =NORMSINV(1-(*p*-value)/2). A conversion spreadsheet is available from: http://osf.io/uygrs/files/

Whenever the observed effect size is lower than expected, it is also possible to change the final sample size on the basis of the effect size observed in an interim analysis. In medical sciences, such *adaptive designs* are controversial (see Chow & Chang, 2008), because they can lead to a statistically significant result without any practical significance, and thus, a new treatment can claim to be "better" without having a practical benefit. Therefore, Proschan et al. (2006) conclude that "Sample size methods based on the treatment effect are like antidepressants: it is best not to need them, but if you do, they work reasonably well." In psychology, there is also a risk to publish statistically significant but practically meaningless effects, but the lack of accurate effect size estimates (e.g., due to publication bias) is a much greater problem. This makes adaptive designs useful for researchers who want to perform studies with sufficient statistical power when effect sizes turn out to be lower than expected based on the published literature, while controlling Type 1 error levels.

When a study is designed with the goal to observe a statistically significant effect, researchers need to perform an *a priori* power analysis, which requires an estimate of the expected effect size. In sequential analyses, this is referred to as the *unconditional power* of a study. If researchers use an adaptive design, interim analyses allow researchers to calculate *conditional power*, which is the conditional probability of a statistically significant benefit at the end of a study, *given the data collected so far*. After an interim analysis, different predictions can be made about the likelihood of observing a statistically significant effect at the end of the trial by assuming that the trend observed in the data collected so far continues or by assuming that the remaining data will yield an effect size equal to the estimated effect size used in the unconditional power analysis. Researchers should keep in mind that effect sizes calculated from small sample sizes have large CIs, and conditional power analyses based on small samples should not be given too much weight. A recent simulation by Schönbrodt and Perugini (2013) provides useful recommendations for the required sample size to yield reasonably stable effect sizes estimates that do not change considerably when additional participants are collected. They suggest a minimum sample size of 55 in each condition for the average effect size in psychology ($d = 0.43$). Researchers thus might want to use these recommendations as a guide to determine when to perform a conditional power analysis.

Researchers should decide whether they want to use an adaptive design, or whether they want to plan for a large sample size, and rely on interim analyses to stop the data collection when a convincing level of significance has been observed (for a discussion, see Tsiatis & Mehta, 2003). A useful recommendation is to use sequential analyses based on a reasonable effect size estimate, but include the possibility of an adaptive design if the effect size is lower, but still of interest to the researcher. Instead of conditional power, researchers can also use *predictive power*, a Bayesian alternative to conditional power (Spiegelhalter, Freedman, & Blackburn, 1986).

DEFINING THE SMALLEST EFFECT SIZE OF INTEREST

If researchers want to terminate the data collection early when an effect size estimate is smaller than a minimum value, it is

important to define what the *smallest effect size of interest* (SESOI) is. In applied research, practical limitations of the SESOI can often be determined on the basis of a cost–benefit analysis. For example, if an intervention costs more money than it saves, the effect size is too small to be of practical significance. In theoretical research, the SESOI might be determined by a theoretical model that is detailed enough to make falsifiable predictions about the hypothesized size of effects. Such theoretical models are rare, and therefore, researchers often state that they are interested in any effect size that is reliably different from zero. Even so, because you can only reliably examine an effect that your study is adequately powered to observe, researchers are always limited by the practical limitation of the number of participants that are willing to participate in their experiment or the number of observations they have the resources to collect.

Whether your choice for a smallest effect size of interest is based on a real-life effect that is desirable, the effect size predicted by a theoretical model, or simply based on practical limitations with respect to the number of participants you can or are willing to collect, it is always possible to define a smallest effect size of interest. For example, I would personally, and in general, consider collecting 500 participants in individual sessions too much effort to be worthwhile. As a consequence, assuming a lower limit of statistical power of 0.80 (or 0.90), the SESOI would be Cohen's $d_s = 0.25$ (or Cohen's $d_s = 0.29$) for a between-subjects *t*-test with an alpha of .05. The SESOI for a paired-samples *t*-test would be Cohen's $d_z = 0.13$ (or Cohen's $d_z = 0.15$). It is useful to move beyond the rather hollow statement that researchers are *in theory* interested in any effect that is reliably different from zero because it allows a researcher to stop the data collection because of futility whenever an effect size is smaller than the SESOI.

In some designs, for example, with a dichotomous dependent variable, it is possible that the outcome of the experiment will yield a significant difference, regardless of the remaining data that a researcher has planned to collect. In these situations, a study can be stopped because the statistical significance of the outcome is completely determined. This is referred to as *curtailment*. In other situations, the observed effect size might be too small to be reliably observed given the maximum number of observations a researcher is willing or able to collect. Lan and Trost (1997) suggest that researchers stop for futility when the conditional power drops below a lower limit, continue with the study when the conditional power is above an upper limit, and extend the study whenever it lies between the lower and upper limits (i.e., increase the planned sample size) to achieve a conditional power that equals the upper limit.

When such an approach is used for replication studies, researchers can prevent inconclusive outcomes in situations where effect sizes might be reliably different from zero but the planned sample size was too small to yield a statistically significant effect. As such, sequential analyses might prove to be a preferable alternative to recent suggestions to perform replication studies using 2.5 times the sample size of the original experiment (Simonsohn, 2014), which could still result in an inconclusive outcome. Sequential analyses, on the other hand, can be continued until a statistically significant result has been observed, the effect size estimate is reliably lower than a minimum value, or the 95% CI around the effect size has a desired width that still includes zero. Although sequential analyses lead to more conclusive outcomes, they can also put a high burden on researchers who perform replications because they need to collect large sample sizes (e.g., Donnellan, Lucas, & Cesario,).

DISCUSSION

When researchers perform studies, they should aim to spend the time of participants and the money they receive from taxes as efficiently as possible. Given the increasing awareness that well-powered studies in psychological research require larger sample sizes than researchers were accustomed to, the general uncertainty around effect size estimates in most research domains, and the risk of inflated Type 1 error rates when results are analyzed repeatedly, an important question is how well-powered experiments should be designed. To do research efficiently, researchers should collect no more data than needed, either because the collected data are sufficiently convincing that the predicted effect exists or because it is unlikely that continued data collection will yield the predicted effect.

This article discussed procedures developed in medical sciences that allow researchers to perform interim analyses while the data collection is still in progress, without inflating the Type 1 error rate. These interim analyses can be used to plan adaptive designs, where conditional power analyses are used to determine the final sample size of a study. The flexibility provided by these statistical procedures makes it possible to collect larger samples more efficiently and reduce the risk of performing studies where the outcome remains inconclusive because of a lack of statistical power. Researchers should determine the rules they rely on to terminate the data collection or extend the sample size in advance, and preferably pre-register these stopping rules such that they can be presented during peer review. Researchers interested in performing studies that use adaptive designs where conditions are dropped, or where the main dependent variable is changed, might want to involve a statistical consultant who guides the researcher through the design process and helps to interpret the outcome of the analyses.

SOME REMARKS ON NULL HYPOTHESIS SIGNIFICANCE TESTING

Although Type 1 errors are the alpha of statistical inferences, they are not its alpha and omega. The goal of research is not to control Type 1 error rates but to discover what is likely to be true. This article discussed procedures how to control Type 1 error rates and plan the size of a sample on the basis of the likelihood that a statistically significant finding will be observed when the data collection is terminated. Throughout this discussion, I hope researchers will remember that inferences from data are only one aspect of the empirical cycle, and good theory and methodology are equally essential for scientific progress.

Null hypothesis significance testing remains a strong tradition in psychological research, but it is only one approach that can be used to draw conclusions from studies. The reliance on statistical significance as a guide to interpret the outcomes of studies, and more specifically its importance in deciding whether results should be published or not, is a major shortcoming of current practices in some research domains within psychology. There are situations where statistical significance addresses the question a researcher is interested in, such as when a carefully controlled experiment aims to examine the novel theoretical prediction that an effect exists. However, researchers are also interested in how likely it is a finding will replicate, which is not easily predicted within a null-hypothesis significance testing (NHST) framework (Miller & Schwarz, 2011). Another important goal of research is to provide an accurate estimation of the value of parameters such as the effect size, which is not the same as observing a statically significant finding (Cumming, 2008; Lakens & Evers, in press).

Using sequential analyses will inevitably lead toward a more Bayesian approach of thinking about what to believe in. Sequential analyses have been considered the "front line" in the debate between Bayesian and frequentist approaches to statistical inference (Spiegelhalter, Abrams, & Myer, 2004), and researchers who embrace sequential analyses will slowly ease into the idea of updating the probability that a hypothesis is true as additional evidence is acquired, which will be a positive development in the long run. Because most researchers use significance tests, and not Bayesian statistics, the goal of this article was to explain how sequential analyses can be used to improve current statistical practices.

Previous articles in which sequential analyses were discussed in light of the fundamental shortcomings of *p*-values gave the impression that sequential analyses procedures are themselves deeply flawed (Wagenmakers, 2007). However, there is no reason why the sequential analyses procedures that have been developed in medical sciences should not be useful for psychological research. Significance tests in sequential analyses will generally lead to similar conclusions about the evidence in favor of the alternative hypothesis as Bayesian analyses would have yielded, and a Bayesian approach to evaluate accumulating data while an experiment is in progress will in practice lead to boundaries to terminate the experiment that lie between the Pocock and O'Brien-Fleming boundaries, if a skeptical prior is chosen with a mean effect of 0 (for a detailed discussion, see Proschan et al., 2006). A Bayesian approach could differ from the analyses within an NHST framework when a prior is chosen that puts a high probability on the alternative hypothesis. However, because the choice of the prior will lead to a different conclusion about the same data, anything else than a skeptical prior is controversial. Besides this difficulty, Bayesian analyses have additional benefits, in that they are more flexible, and allow researchers to perform interim analyses after every observation, which in an NHST framework would lead to high costs in terms of error rates or statistical power.

DATA MONITORING AND EXPERIMENTER BIAS

In large medical trials, tasks such as data collection and statistical analysis are often assigned to different individuals, and it is considered good practice to have a data and safety monitoring board that is involved in planning the experiment and overseeing any interim analyses. In psychology, such a division of labor is rare, and it is much more common that researchers work in isolation. The number of times data are analyzed matters for the Type 1 error level in an NHST framework, and the planned final sample size matters for the adjusted *p*-value, effect size, and 95% CIs calculated when the data collection is terminated. As a consequence, it is essential that sequential analyses are reported when they are performed.

The best approach to guarantee this is to pre-register the study design. Researchers should specify the planned sample size (or in adaptive designs, the criteria on which the final sample size will be determined), the number of interim analyses, and the rules that will be used to terminate the data collection, in addition to the number of different conditions, dependent variables, and other relevant aspects of the study. Whenever an experiment is submitted for publication, researchers can link to the pre-registration document, and reviewers can check whether the data collection and analysis adhered to the predefined plan (see Nosek & Lakens, in press). Note that although pre-registration is important for science, there is often no immediate benefit for the individual researcher to pre-register studies. Because sequential analyses do offer an immediate benefit (such as reducing the required sample size for well-powered studies to at least 20–30%), and sequential analyses should be pre-registered, embracing sequential analyses might also increase the prevalence of pre-registered experiments.

As an example of a pre-registered sequential analysis, researchers could state:

> Based on an expected effect size of Cohen's $d = 0.43$, a power analysis indicated that for a two-sided test with an alpha of .05, a desired statistical power of .9, and two looks using a linear spending function, a total of 244 participants are needed. If the expected difference is significant at the first interim analysis (after 100 participants or time = .41, with an alpha boundary of .0205) the data collection will be terminated. The data collection will also be terminated when the observed effect size is smaller than the SESOI, which is set at $d = 0.29$ based on the researcher's willingness to collect at most 400 participants for this study, and the fact that with one interim analysis 400 participants provide .8 power to detect an effect of $d = 0.29$. If the interim analysis reveals an effect size larger than 0.43 (while $p > .0205$), the data collection will be continued until 244 participants have been collected. If the effect size lies between the SESOI ($d = 0.29$) and the expected effect size ($d = 0.43$), the planned sample size will be increased based on a conditional power analysis to achieve a power of .9 (or to a maximum of 400 participants in total). The second analysis is performed at an alpha boundary of .0358.

A benefit of sequential and adaptive designs is that they allow much flexibility, while controlling the level of Type 1 errors, and researchers can pre-register much more flexible designs. After an interim analysis, researchers can decide to terminate an experiment, increase the sample size on the basis of a conditional power analysis performed when enough participants have been collected to yield a stable effect size estimate (Schönbrodt & Perugini, 2013), or even stop

collecting data for specific conditions (referred to as *drop-the-losers* designs, see Chow & Chang, 2008) while continuing data collection for the remaining conditions (Bauer & Köhne, 1994), all while keeping the Type 1 error below a desired level. Note that the current discussion of sequential analyses assumes that a single dependent variable is analyzed in a confirmatory fashion. If researchers want to examine multiple dependent variables, Bonferroni corrections can be applied. Furthermore, exploratory analyses remain possible and could prove interesting, although because of the unknown increase in alpha level, such findings are preferably replicated before taken too seriously.

One might feel that Type 1 error control is not necessary, and that everything is allowed, as long as people clearly report what they have performed. This idea assumes that reviewers can decide whether the results of a study provide convincing support for a hypothesis or not. However, because people are notoriously bad at thinking about probabilities, and specifically bad at thinking about conditional probabilities such as *p*-values, letting reviewers decide whether a chosen procedure leads to a result that is convincing might be asking for trouble. Instead, an approach where researchers clearly specify the flexibility they desire in advance, and control the overall alpha level, is a more objective procedure to determine how convincing the data are.

Experimenter bias is important to consider when performing a study under normal circumstances (e.g., Klein et al., 2012) but becomes even more important to consider when the experimenter has performed an interim analysis. The risk of subtle changes in procedures or differences in the way participants are treated may increase if the experimenter has knowledge of the effects the manipulation has had on previous participants. The experimenter needs to be blind to conditions and ideally to the hypothesis. In situations where a researcher is interested in the difference between two groups regardless of the direction of this difference (i.e., the study is deemed worthwhile regardless of whether $X_1 > X_2$ or $X_1 < X_2$), the researcher might ask someone else to perform the interim analysis, such that when the outcome of the interim analysis indicates that data collection should be continued until a specific number of participants has been collected, the experimenter remains blind to the direction of the effect.

CONCLUSION

Sequential analyses provide potentially important benefits for psychological science, because they allow researchers to perform well-powered studies more efficiently. It is surprising that the statistical procedures discussed here have not been used, or even considered, by psychologists, especially in light of recent work on ways to improve the reliability of psychological research. I hope this brief introduction, together with the supplementary materials, will provide researchers interested in sequential analyses with an easy-to-follow explanation of how these procedures can be incorporated into their research.

REFERENCES

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

Bauer, P., & Köhne, K. (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics*, *50*, 1029–1041.

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

Chow, S. C., & Chang, M. (2008). Adaptive design methods in clinical trials —A review. *Orphanet Journal of Rare Diseases*, *3:11*. DOI: 10.1186/1750-1172-3-11

Chow, S. C., Shao, J., & Wang, H. (2003). *Sample size calculations in clinical research*. New York: Marcel Dekker.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York, NY: Routledge Academic.

Cumming, G. (2008). Replication and *p* intervals: *p* values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science*, *3*(4), 286–300.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge.

Dodge, H. F., & Romig, H. G. (1929). A method of sampling inspection. *Bell System Technical Journal*, *8*(4), 613–631.

Donnellan, M. B., Lucas, R. E., & Cesario, J. (in press). On the association between loneliness and bathing habits: Nine replications of Bargh and Shalev (2012) study 1. *Emotion*.

Ellemers, N. (2013). Connecting the dots: Mobilizing theory to reveal the big picture in social psychology (and why we should do this). *European Journal of Social Psychology*, *43*, 1–8. DOI: 10.1002/ejsp.1932

European Medicines Agency. (2006). Point to consider on methodological issues in confirmatory clinical trials with flexible design and analysis plan. The European Agency for the Evaluation of Medicinal Products Evaluation of Medicines for Human Use. CPMP/EWP/2459/02, London, UK.

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

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

Klein, O., Doyen, S., Leys, C., Miller, S., Questienne, L., & Cleeremans, A. (2012). Low hopes, high expectations: Expectancy effects and the replicability of behavioral experiments. *Perspectives on Psychological Science*, *7*(6), 572–584.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for *t*-tests and ANOVAs. *Frontiers in Psychology*, *4:863*. DOI: 10.3389/fpsyg.2013.00863

Lakens, D., & Evers, E. (in press). Sailing from the seas of chaos into the corridor of stability: Practical recommendations to increase the informational value of studies. *Perspectives on Psychological Science*.

Lan, K. G., & DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika*, *70*(3), 659–663.

Lan, K. K. G., & Trost, D. C. (1997). Estimation of parameters and sample size re-estimation. In *Proceedings-Biopharmaceutical Section American Statistical Association*. American Statistical Association, 48–51.

Leifer, E. S., & Geller, N. L. (2012). Monitoring randomized clinical trials. In K. Hinkelmann (Ed.), *Design and analysis of experiments, special designs and applications* (vol. *3*, pp. 213–249). Hoboken, NJ: Wiley.

Miller, J., & Schwarz, W. (2011). Aggregate and individual replication probability within an explicit model of the research process. *Psychological Methods*, *16*, 337–360. DOI: 10.1037/a0023347

Murayama, K., Pekrun, R., & Fiedler, K. (2013). Research practices that can prevent an inflation of false-positive rates. *Personality and Social Psychology Review*. DOI: 10.1177/1088868313496330

Nosek, B. A., & Lakens, D. (in press). Registered reports: A method to increase the credibility of published results. *Social Psychology*. DOI: 10.1027/1864-9335/a000192

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, *64*(2), 191–199.

Pocock, S. J. (1992). When to stop a clinical trial. *BMJ [British Medical Journal]*, *30*, 235–240.

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

Reboussin, D. M., DeMets, D. L., Kim, K., & Lan, K. K. (2000). Computations for group sequential boundaries using the Lan-DeMets spending function method. *Controlled Clinical Trials*, *21*, 190–207.

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, *47*, 609–612. DOI: 10.1016/j.jrp.2013.05.009

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

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2013). Life after *p*-hacking. Meeting of the Society for Personality and Social Psychology, New Orleans, LA, 17-19 January 2013. Available at http://dx.doi.org/10.2139/ssrn.2205186

Simonsohn, U. (2014). Evaluating replication results. Downloaded on September 30th 2013 from http://ssrn.com/abstract=2259879. DOI: 10.2139/ssrn.2259879

Spiegelhalter D., Abrams K., & Myer J. (2004). *Bayesian approaches to clinical trials and health-care evaluations*. Hoboken, NJ: Wiley.

Spiegelhalter, D. J., Freedman, L. S., & Blackburn, P. R. (1986). Monitoring clinical trials: Conditional or predictive power? *Controlled Clinical Trials*, *7*, 8–17.

Tsiatis, A. A., & Mehta, C. (2003). On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika*, *90*, 367–378.

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

Wittes, J., & Brittain, E. (1990). The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine*, *9*, 65–72.