General Article

Advances in Methods and Practices in Psychological Science
January-March 2024, Vol. 7, No. 1, pp. 1–20

DOI: 10.1177/25152459231212128
www.psychologicalscience.org/AMPPS

How to Safely Reassess Variability and Adapt Sample Size? A Primer for the Independent Samples *t* Test

Lara Vankelecom, Tom Loeys, and Beatrijs Moerkerke
Department of Data-Analysis, Ghent University, Ghent, Belgium

Abstract

When researchers aim to test hypotheses, setting up adequately powered studies is crucial to avoid missing important effects and to increase the probability that published significant effects reflect true effects. Without a priori good knowledge about the population effect size and variability, power analyses may underestimate the true required sample size. However, a specific type of a two-stage adaptive design in which the sample size can be reestimated during the data collection might partially mitigate the problem. In the design proposed in this article, the variability of the data collected at the first stage is estimated and then used to reassess the originally planned sample size of the study while the unstandardized effect size is fixed at a smallest effect size of interest. In this article, we explain how to implement such a two-stage sample-size reestimation design in the setting in which interest lies in comparing means of two independent groups. We investigate through simulation the implications on the Type I error rate (T1ER) of the final independent samples *t* test. Inflation can be substantial when the interim variance estimate is based on a small sample. However, the T1ER approaches the nominal level when more first-stage data are collected. An R-function is provided that enables researchers to calculate for their specific study (a) the maximum T1ER inflation and (b) the adjusted α level to be used in the final *t* test to correct for the inflation. Finally, the desired property of this design to better ensure the power of the study is verified.

Keywords

null hypothesis significance test, power analysis, sample-size reestimation, internal pilot, adaptive design

Received 10/6/22; Revision accepted 10/9/23

Performing a null hypothesis significance test (NHST) on data can result in erroneous conclusions by either incorrectly rejecting the null hypothesis or incorrectly failing to reject the null hypothesis (Neyman & Pearson, 1928). The former is referred to as a "Type I error." The maximum Type I error rate (T1ER; i.e., α) of a study is controlled, under certain statistical assumptions, at the significance level used for the NHST (typically 5%). The latter is referred to as a "Type II error." The Type II error rate (i.e., β) of a study received less attention in the past. However, researchers increasingly realize the importance of also controlling this error rate, which can be achieved by properly powering the study (i.e., $1-\beta$; e.g., aiming for a power of 80% limits the Type II error rate to 20%). Power is the probability to detect an effect, when there actually is one, and depends on, among other things, the collected sample size. Thus, careful consideration of the sample size is crucial to ensure that a study has sufficient statistical power. Unfortunately, sample-size decisions are in practice still often based on resource constraints (e.g., time, money, or unavailability of data), a general rule of thumb or common practice that dominates the scientific field, or intuition (Bakker et al., 2016; Bakker & Wicherts, 2011; Lakens, 2022). These approaches generally lead to feasible small-sample sizes that lack the power to detect effects. The power of a typical psychological study investigating the

Corresponding Author:
Lara Vankelecom, Department of Data-Analysis, Ghent University, Ghent, Belgium
Email: lara.vankelecom@ugent.be

difference between two independent groups has been estimated to be below 50% (Cohen, 1990) or even 35% (Bakker et al., 2012).

The importance of proper statistical power in research cannot be overstated. Many of the problems currently confronting the field of psychology partly originate from underpowered research (e.g., file-drawer problem, replication crisis; Fraley & Vazire, 2014; Rosenthal, 1979). Studies with low power are less likely to detect an existing effect and therefore have a higher probability to turn out nonsignificant. Nonsignificant research results are hard to publish in scientific journals (i.e., publication bias) and most likely just disappear in the researcher's file cabinet (Fanelli, 2012). Such research is inefficient, wasteful, and above all, uninformative. Failing to detect an effect leaves the researcher in doubt: Did the negative outcome follow because no true effect exists or simply because the sample size of the study was too small to detect it? In addition, when low-powered research does achieve significance and gets published, several problems arise. First, the reported effect-size estimates of these studies tend to be upwardly biased. In studies with small sample sizes (and hence, low power), the effect-size estimate will vary widely around the true value. Significant results rise only from large effect sizes, leading to inflated estimates being reported in the literature. Second, the findings of these studies are less successfully replicated because they are more likely to reflect a false-positive result. Researchers who initially obtained nonsignificance for their primary analysis (which often occurs when faced with low power) may resort more easily to questionable research practices, such as p-hacking, to manipulate the "unlucky" result into a significant one that can be published (Fanelli, 2009; John et al., 2012). These practices often lead to an inflation of the T1ER, which increases the chance that the published result is a false-positive one. The proportion of false positives in a research area is determined not solely by the possibly inflated significance level used in the individual studies but also by the overall statistical power of the studies in that area (Ioannidis, 2005; Pashler & Harris, 2012; Wacholder et al., 2004). In research areas with low typical study power, the false-positive rate (i.e., the number of false positives divided by the sum of false positives and correct hits) will be high because there will be fewer correct hits compared with well-powered research areas. Thus, when a study reports a significant effect, the probability that it reflects a false-positive result is higher for an underpowered study (area) than for a properly powered study (area).

Adequately powered research is crucial to tackle the problems described above. To try to ensure that an empirical study, in which the goal is to test a specific hypothesis, is appropriately powered, the sample size of the study is ideally determined by performing a sample-size calculation (also called a "power analysis"). Sample-size calculation gives an indication on the minimum number of observations that need to be collected for a study to find a certain effect with sufficient power. In the next section, we explain how such sample-size calculation can be performed for the two-sample *t* test.[1] In this article, we focus on sample-size calculations in which the effect (i.e., difference in means) and the variance of the data are separately defined as opposed to using a standardized effect. In addition, we recommend to fix the effect at a theoretical meaningful effect, which the researcher can define based on their own judgment and research goals. This leaves researchers with the estimation of the variance before sample-size calculation can be performed. How such variance estimate can be obtained depends on how the sample-size calculation is implemented into the study design. In this article, we discuss two possible approaches. The first option, which is also the most dominant and well-known approach in psychological research, is to perform the sample-size calculation a priori. This approach fits within the traditional fixed design, in which the study design (including the sample size) needs to be fixed before data collection starts. Calculating the required sample size beforehand ensures strict control of the T1ER (when no questionable research practices are conducted) but is difficult because the population variance needs to be estimated a priori without any study data. As we show, this approach will generally not solve the problem of underpowered studies. A second option, which is much less known in psychological research and is mainly used within clinical trials, is to allow the sample-size calculation to be conducted sometime during the course of the study, thereby enabling to use the first collected data to estimate the variance parameter necessary for sample-size calculation. This design is referred to as the "internal-pilot-study" (IPS) design and belongs to the group of adaptive designs because the sample size is not fixed in advance. Note that the effect (i.e., difference in means) is not estimated on the first collected data but remains set at the smallest meaningful effect during the midstudy sample-size calculation. As we show, using the IPS design to determine the sample size of a study can preserve the T1ER with some (minor) correction and might improve the chance of obtaining a well-powered study compared with the a priori power analysis approach because a more accurate estimate of the variance can be obtained (explained later in article). We conclude by providing a detailed step-by-step guide that explains the implementation of the IPS design in a study employing the two-sample *t* test, including a practical example.

Sample-Size Calculation for the Two-Sample *t* Test

Formula for Student t *test*

Suppose a researcher is interested in comparing the means of two groups, assuming the observations in each group are obtained from a normal distribution with each a population mean μ_j ($j = 1, 2$) and an unknown but common variance σ^2. The researcher can then test the null hypothesis, stating there is no difference in means ($H_0: \mu_1 - \mu_2 = 0$) against the two-sided alternative ($H_1: \mu_1 - \mu_2 \neq 0$), with the Student t test.[2] For a given α, the researcher wants the test to have a power of $1 - \beta$ to detect a certain difference in means (δ; see later). The necessary sample size for uncovering δ with this specified power can be calculated from the following:

$$N_1 = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{\delta^2 \times \pi_2} \times \sigma^2 \times (1 + \pi_2), \quad (1)$$

where z_γ denotes the γ-percentile of the standard normal distribution[3] and π_2 signifies the desired ratio of the sample size of Group 2 to Group 1 ($\pi_2: N_2/N_1$). Although it is generally recommended to collect two equally sized groups whenever possible, the inclusion of the π_2 parameter in the equation enables sample-size calculation in situations in which gathering equally sized groups proves challenging. For instance, when data for a control group (Group 1) are much more readily available than data for a treatment group (Group 2), the researcher may decide to make Group 1 twice as large as Group 2 by setting π_2 to .5. Whereas the required sample size of the study for Group 1 is directly given by Equation 1, the required sample size for Group 2 can easily be obtained by multiplying N_1 with π_2. Because the common population variance σ^2 is generally unknown in practice, it first needs to be estimated before the sample-size calculation can be performed. When replacing the common population variance σ^2 by its estimate, the resulting sample sizes from Equation 1 should be viewed as estimates ($\hat{N}_1$ and $\hat{N}_2$) of the true unknown required sample sizes (N_1 and N_2, respectively).

Note that the sample-size calculation can also be performed with a standardized effect size, in which the difference in means (δ) and the variability of the data (σ) are combined in one measure (e.g., Cohen's *d*). In practice, researchers often power their study for a small, medium, or large standardized effect. This approach is tempting because the researcher can just use the rules of thumb by Cohen (1988) to specify the standardized effect for the sample-size formula and does not need to worry about estimating the variance. However, the challenge with this approach is that it is difficult to disentangle δ and σ and the researcher does not know what difference in means the study is actually powered for; it could be a very small or very large difference, depending on the actual variance (which is not known). As a result, the researcher does not have great control over or insights into the statistical power of the study (Cummings, 2011). In this article, we therefore focus only on the best practice of performing a sample-size calculation with a separate estimate for the effect δ and the variance σ^2.

Formula for Welch t *test*

Consider the same scenario as above, however now, the researcher does not assume a common variance between both groups. Then, the Welch t test[4] should be performed to test for significance, and the formula for sample-size calculation has to be modified to allow for unequal population variances between groups:

$$N_1 = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2}{\delta^2 \times \pi_2} \times \sigma_1^2 \times (\pi_1 + \pi_2), \quad (2)$$

where the additional parameter π_1 represents the ratio of the population variance of Group 2 relative to Group 1 ($\pi_1: \sigma_2^2/\sigma_1^2$). Again, the population variances of both groups first needs to be estimated to be able to estimate π_1 and perform the sample-size calculation to obtain $\hat{N}_1$ and $\hat{N}_2$.

Sample-size calculation for the independent samples t test like in Equation 1 and Equation 2 can also be performed with the R function `power_t_test()` from the package *MESS* (Ekstrøm, 2020), which yields an exact (as opposed to an asymptotic approximate) result of the required sample size.

Defining δ

Because the purpose of an independent samples t test is to uncover the effect that is present in the population, δ in Equations 1 and 2 is typically defined as an estimate of this population effect. However, obtaining an accurate estimate of the population effect is very challenging, especially in the fixed-sample design, in which it needs to be defined before any data of the study are collected. In the first place, it is unsure beforehand whether a population effect exists, after all. In the second place, techniques such as using estimates of the population effect published in related literature (e.g., a meta-analysis or a previous similar study; Fritz et al., 2012; Polit & Beck, 2004) for the sample-size calculation of the current study generally lead to underpowered studies. Because of publication bias, reported estimates are inflated (i.e., larger than the true effect) and will cause

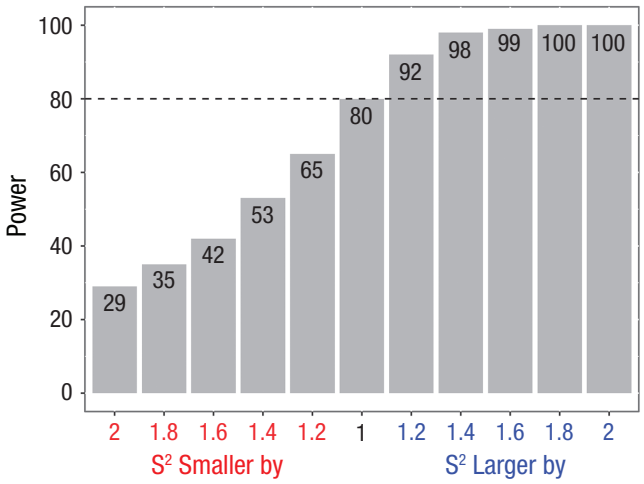

Fig. 1. The empirical power of a study when the a priori power analysis is performed with a δ of 0.5, α level of 5%, desired power level of 80%, and a variance estimate ranging from 2 times smaller to 2 times larger than the true common variance (when the true difference in means equals δ).

the calculated sample size of the current study to be lower than what is needed for finding the population effect with specified power (Albers & Lakens, 2018; Brysbaert, 2019; Lakens, 2022). In addition, effects in the literature are often reported in the standardized way, making them useless for obtaining an estimate of the difference in population means δ.

Instead of trying to estimate the population effect, the researcher could just perform the sample-size calculation for a meaningful effect. A meaningful effect can be defined as the smallest effect that the researcher finds worthwhile to discover with the study and is in the literature often referred to as the "smallest effect size of interest" (SESOI). When one uses an (unstandardized) SESOI for δ in Equations 1 and 2, the resulting sample size will ensure that a population difference in means of this size or larger is detected with desired power. When the true difference is smaller than the SESOI, and therefore too small to be considered interesting according to the researcher, it can still be identified with the collected sample, however, only with (much) lower power. For example, when a researcher decides, after careful consideration, to conduct the sample-size calculation with a SESOI of 0.5, only a true difference in means of 0.5 units or higher will be detected with specified power. The advantage of this approach is that the researcher can be confident the smallest difference in means considered interesting is found by the *t* test with desired power (however, this is the case only when the variance is also accurately estimated; see further) and therefore can draw proper conclusions about the possible absence of an effect (Aberson, 2019; Albers & Lakens, 2018; Lakens, 2022; Lenth, 2001). What SESOI should be considered for the study is up to the researcher but should optimally be based on theoretical predictions or practical considerations. For a review of different approaches to determine the SESOI, see Lakens et al. (2018). When fixing δ at a theoretical smallest meaningful effect, estimating the variance of the data is the only remaining challenge before sample-size calculation can be performed. In the next two sections, we outline two different strategies to obtain such variance estimate by implementing the sample-size calculation either a priori or midstudy.

A Priori Sample-Size Calculation

The most common way in research to perform sample-size calculation for a study is a priori because this is in line with the dominant fixed-sample approach. However, this also means that an estimate of the population variance(s) needs to be available before any data are collected. Typically, very little information is available in published literature about the variance. Even when a variance estimate or the data of a related study are provided, the variance may differ from that of the current study, for example because of subtle differences in experimental setup or because of a more or less heterogeneous population. Therefore, one of the only options to obtain an a priori variance estimate is to gather a pilot sample for both groups before the main study and use these data to estimate either the common variance σ^2 (for sample-size calculation for the Student *t* test; Equation 1) or the variance ratio π_1 (for sample-size calculation for the Welch *t* test; Equation 2). Because pilot data are supposed to be discarded after variance estimation, the researcher typically limits the size of the pilot sample to avoid waste of data (e.g., only five observations per group). Because of this small sample, the obtained estimates for σ^2 or π_1 will vary substantially around their true values. Figure 1 illustrates what happens to the power of a study to detect an unstandardized SESOI of 0.5 (when this effects truly exists in the population) when an estimate that is either smaller or larger than the true common variance is used for σ^2 in Equation 1. For instance, when the estimate obtained from the small pilot sample is 2 times smaller than the true variance, the power to detect the SESOI drops from the intended 80% to 29%. Similar consequences on power are present when misspecifying σ_1^2 and π_1 in Equation 2. Although underpowered studies are prevalent, overpowered studies are much less common in research because large variance estimates from the pilot sample that result in excessively large sample sizes are often not pursued for further study (i.e., follow-up bias). To avoid the risk that a study is insufficiently powered to detect a certain mean difference of interest, a more precise estimate of the population

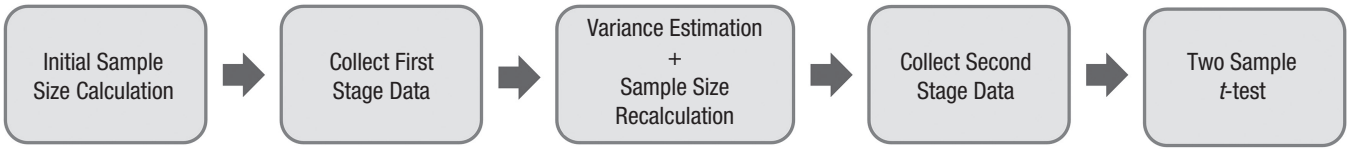

Fig. 2. The different steps in an internal-pilot-study design.

variance(s) is crucial. This can be accomplished only by using more data for the variance estimation. The IPS design, which is outlined in the next section, can realize this without having to waste any pilot data.

IPS Design

In the IPS design, the pilot sample used to estimate the common variance σ^2 or the variance ratio π_1 is "internalized" into the final data and is no longer discarded. In other words, σ^2 or π_1 is estimated on the first collected observations of the study and is then used to reestimate the final sample size with Equation 1 or Equation 2, respectively, when δ is from the beginning fixed at a smallest meaningful mean difference. The IPS design is therefore also often referred to as "blinded" sample-size reestimation ("blinded" because the population mean difference is not estimated interim[5]). The advantage of this design is that more accurate variance estimates can be obtained for sample-size calculation without wasting any data, thereby improving the power of the study. The disadvantage is that a small inflation of the T1ER may occur (see later). The IPS design is an adaptive design because the sample size is not predetermined as usual but can still be modified midstudy. More specifically, it is a two-stage adaptive design because there is only one interim analysis in which the variance is estimated. Figure 2 gives a general overview of the different steps in the IPS design. The first step consists of performing an initial sample-size calculation with an initial variance guess and is solely intended to give the researcher an idea of a reasonable internal pilot size (e.g., 50% of the initially planned sample size is collected as internal pilot). This first step can also be omitted when researchers have already decided beforehand how many internal pilot data they can or want to collect. The size of the internal pilot is completely up to the researcher. However, it is important to keep in mind that only large internal pilots will lead to more precise variance estimates. After gathering these first-stage data, the variance is estimated and used to update the sample size of the study. Then, the data collection proceeds until the reestimated sample size is reached (i.e., second-stage data). At the end, the data of both stages are pooled together for each of the two groups, and the two-sample *t* test is performed in the same way as in the regular fixed design.

The IPS design was first introduced by Stein (1945) and was then further elaborated within the clinical-trials literature by Wittes and Brittain (1990) and Birkett and Day (1994). Since then, numerous research has been conducted on the IPS design for the two-sample *t* test. However, all this research focused exclusively on describing and investigating the design for studies assuming a common variance σ^2 (Student *t* test) and having two equally sized groups. To make the design accessible for all kinds of studies using the two-sample *t* test, we constructed the sample-size formulas in Equations 1 and 2 to use at the interim analysis. Equation 1 can be used for sample-size reestimation by studies aiming for the Student *t* test, and Equation 2 can be used for studies planning to use the Welch *t* test, both also allowing unequal sample sizes between the two groups. Making the sample-size reestimation procedure also accessible for the Welch *t* test is important because there are typically no strong reasons in psychological research to assume a common variance (Delacre et al., 2017). Psychological studies often compare outcomes in preexisting groups, which already have a different variance from the beginning. Even when the group assignment in (quasi-) experimental designs is completely randomized and the variances of the outcome in the two groups are initially the same, deviation from the homoscedasticity assumption may occur later when the experimental treatment is induced. In what follows, the procedure for interim variance estimation will be reviewed in more detail: first for the Student *t* test (estimation of σ^2) and then for the Welch *t* test (estimation of π_1).

For the Student t *test*

When there is a common variance in the data, the σ^2 parameter of Equation 1 can be estimated at the interim analysis in two different ways: blinded or unblinded. Blinded variance estimation implies that the treatment-group information remains hidden at the interim analysis. When it is not known which observations belong to which group, the only option for estimating the population variance is to pool all the first-stage data together (over the two groups) and calculate a naive one-sample variance estimate. However, this naive estimator is biased when the effect in the population is nonzero. More specifically, the naive estimator will then overestimate the true common variance, which leads to unnecessarily

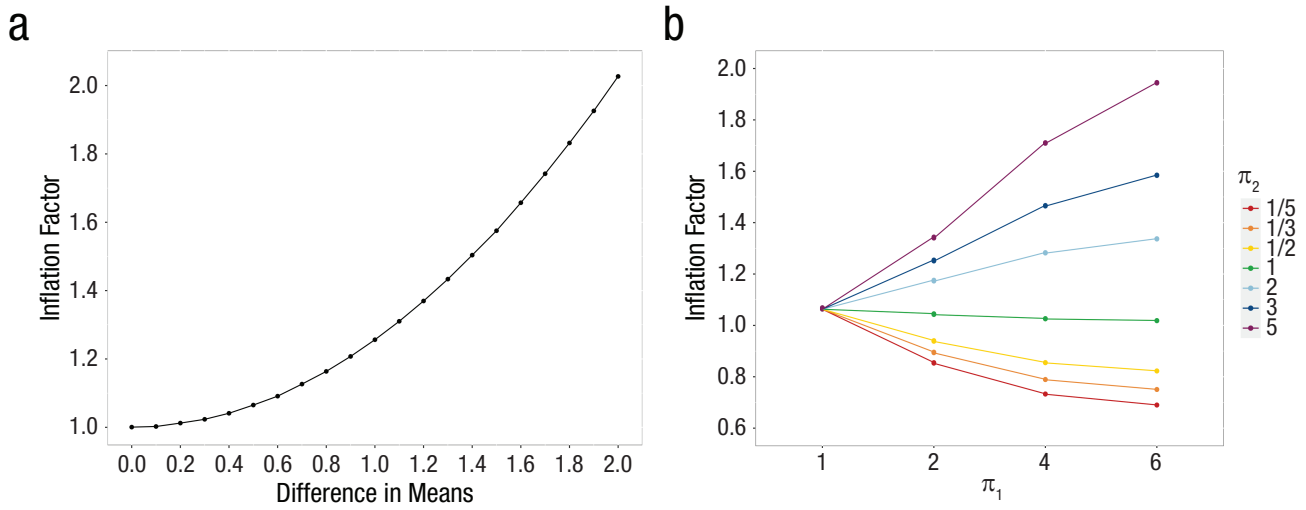

Fig. 3. The bias of the naive estimator in terms of inflation factor (= sample-size estimate with naive estimator relative to sample-size estimate with true variance). (a) The inflation factors when the data of two groups were simulated with a common population variance of 1 and with several values of the true effect (i.e. difference in means δ). (b) The inflation factors when the two groups have an unequal population variance ($\pi_1 \neq 1$) in combination with either equal group sizes ($\pi_2 = 1$) or unequal group sizes ($\pi_2 \neq 1$), the true difference in means is fixed at 0.5, and the internal-pilot groups are equal in size.

large final sample sizes when used as estimator for σ^2 in Equation 1. Figure 3a illustrates how much the true required sample size is overestimated with the naive variance estimator for different effect sizes in the population. The amount of overestimation depends only on the *ratio* of the effect present in the population relative to the variance. For example, when the difference in population means is equally large as the common variance in the data (i.e., $\delta = 1$), the inflation factor is about 1.25, meaning the researcher will, on average, collect 25% more data than necessary. The naive variance estimate obtained at the interim analysis can be refined by adjusting it with the effect under the assumed alternative hypothesis (Gould & Shih, 1992; Zucker et al., 1999). However, this adjusted estimator still overestimates the true variance (and therefore leads to unnecessarily large sample sizes) when the true effect is larger than what is prespecified under the alternative (Friede & Kieser, 2001; Miller, 2005). Despite the risk of collecting more data than necessary, Kieser and Friede (2003) demonstrated that the T1ER of the final Student *t* test is approximately controlled at the nominal α level with blinded interim variance estimation, both with the naive and the adjusted variance estimator.

The alternative is to unblind the treatment groups and use this extra information to calculate the pooled variance estimate of the internal-pilot data. Using the pooled estimator for σ^2 in Equation 1 does not result in a systematical overestimation of the true required sample size but does, however, lead to another problem. With unblinded variance estimation, the pooled variance estimated at the end of the data collection, which is used to construct the final Student *t* statistic, will underestimate, on average, the true common variance (Friede & Kieser, 2006; Wittes & Brittain, 1990; Wittes et al., 1999). The reasoning is as follows. An initial overestimation of the true variance based on the internal-pilot sample may return more easily to the true value at the end of the data collection compared with an initial underestimation of the variance because the former will have a larger amount of additional observations in the second stage (Friede & Miller, 2012). Because of this negative bias in the final pooled variance estimator, an inflation of the T1ER may occur. This inflation will be larger for smaller internal-pilot samples because the average underestimation of the final pooled variance will be larger (more variation of the initial estimates around the true value). With blinded variance estimation (discussed above), the underestimation of the final pooled variance is masked by the unnecessarily large sample sizes that are often collected in the second stage with this approach, resulting in (almost) no T1ER inflation. With this knowledge at hand, it is up to the researcher to decide which procedure to use for interim estimating the common variance. Is it more important to have a strict control of the α level (with the possibility of collecting too much data; blinded estimation) or to have a more accurate sample-size estimation to save resources (however, with a slight possible inflation of the T1ER; unblinded estimation)?

For the Welch t *test*

When population variances are not equal, using the blinded naive estimator might not always be the best

choice. In that case, there is only one variance estimate, and the researcher has no information on π_1 and needs to proceed with Equation 1 instead of Equation 2. Figure 3b shows the bias in the naive estimator when having unequal population variances in combination with different group-size ratios and when the true difference in means equals 0.5. Note that the inflation factor at $\pi_1 = 1$ (which is the same for all group-size ratios) corresponds to the inflation factor for the true difference in means of 0.5 displayed in Figure 3a. For equally sized groups, the inflation factor remains relatively stable when the ratio of population variances moves from equal to unequal. When dealing with equally sized groups, blinded estimation for the Welch *t* test can thus be performed with similar consequences on the inflation of the estimated sample size as displayed in Figure 3a (however, this is the case only when also having equally sized internal-pilot groups!). For unequally sized final groups (with unequal variances), the naive estimator becomes seriously biased: It largely overestimates the true sample size when the largest group has the largest variance ($\pi_2 > 1$) and underestimates the true sample size when the smallest group has the largest variance ($\pi_2 < 1$). These results are even further exacerbated when the true difference in means is larger than 0.5 and/or when the amount of pilot data for the two groups with unequal variances is not the same. Even though a T1ER inflation is avoided with blinded variance estimation, when dealing with unequally sized groups (or unequally sized internal-pilot groups), it is not recommended to perform the interim variance estimation for the Welch *t* test blinded.

The alternative is again to unblind the two treatment groups and estimate the population variance of Group 1 and Group 2 separately. These estimates can then be used to construct $\hat{\pi}_1$, which can be plugged into Equation 2 together with the estimated variance of Group 1 for the sample-size reestimation. We refer to this estimation as the "π_1 estimator." Similarly as for the unblinded pooled estimator for the Student *t* test, the unblinded estimation procedure for the Welch *t* test yields an unbiased estimation of the true required sample size but does, however, come with a slight inflation of the T1ER. The T1ER of the IPS design with unblinded variance estimation is now explored for both the Welch *t* test and the Student *t* test in the next section.

T1ER of the IPS Design

Method

A simulation study was performed to investigate the consequences of using the IPS design with unblinded variance estimation on the T1ER of the final two-sample *t* test when no correction of the α level is implemented. The simulation code is available at https://osf.io/mabrz/. The T1ER of the Student *t* test was calculated for the IPS procedure with the pooled variance estimator, whereas that of the Welch *t* test was calculated for the IPS procedure with the π_1 estimator. To obtain the T1ER levels, the data of both groups were simulated under the null hypothesis of no difference in means. However, a specific value of the smallest difference in means considered interesting had to be specified to perform the midcourse sample-size calculation (because the difference in means is not estimated based on the first-stage data). This value of the SESOI was chosen in such a way that in combination with the variance with which the data of both groups were simulated, the resulting true required sample size (to detect the SESOI with 80% power) was larger than the internal-pilot size by a certain multiplication factor λ.

The T1ER was explored for different scenarios: either two equally sized ($\pi_2 = 1$) or two unequally sized ($\pi_2 \neq 1$) treatment groups in combination with either an equal variance ($\pi_1 = 1$) or an unequal variance ($\pi_1 \neq 1$) between them. The simulation study was initially carried out with equal internal-pilot sizes for the two groups even when the final group sizes were unequal ($\pi_2 \neq 1$). However, it is not a realistic assumption that researchers will always be able to collect the same amount of first-stage data for both groups. To gain some insights in how unequal internal-pilot sizes affect the T1ER, the simulation study was performed again with internal-pilot sizes that followed the same π_2 ratio as the final sample sizes (while keeping the total internal-pilot size identical to the simulation with equally sized pilots). Further details on the simulation study can be found in Appendix A. The T1ER results are first discussed when the correct IPS procedure is applied in the correct scenario, that is, the procedure with pooled estimator and Student *t* test when homoscedasticity holds (a common population variance) and the procedure with the π_1 estimator and Welch *t* test when homoscedasticity does not hold (unequal population variances). At the end of this section, we show the impact on the T1ER of both procedures when homoscedasticity/heteroscedasticity was inaccurately assumed.

Results: equal internal-pilot size

In Figure 4, the T1ER of the Student *t* test in the IPS procedure with the pooled estimator is displayed for the scenarios in which the two groups have an equal variance (Fig. 4a), and the T1ER of the Welch *t* test in the IPS procedure with the π_1 estimator is displayed for the scenarios with unequal variances between groups (Fig. 4b). The Student T1ER does not seem to be influenced by how unequal the group sizes of the study are. The Welch T1ER, on the other hand, slightly increases for more unequal variances/group sizes when the

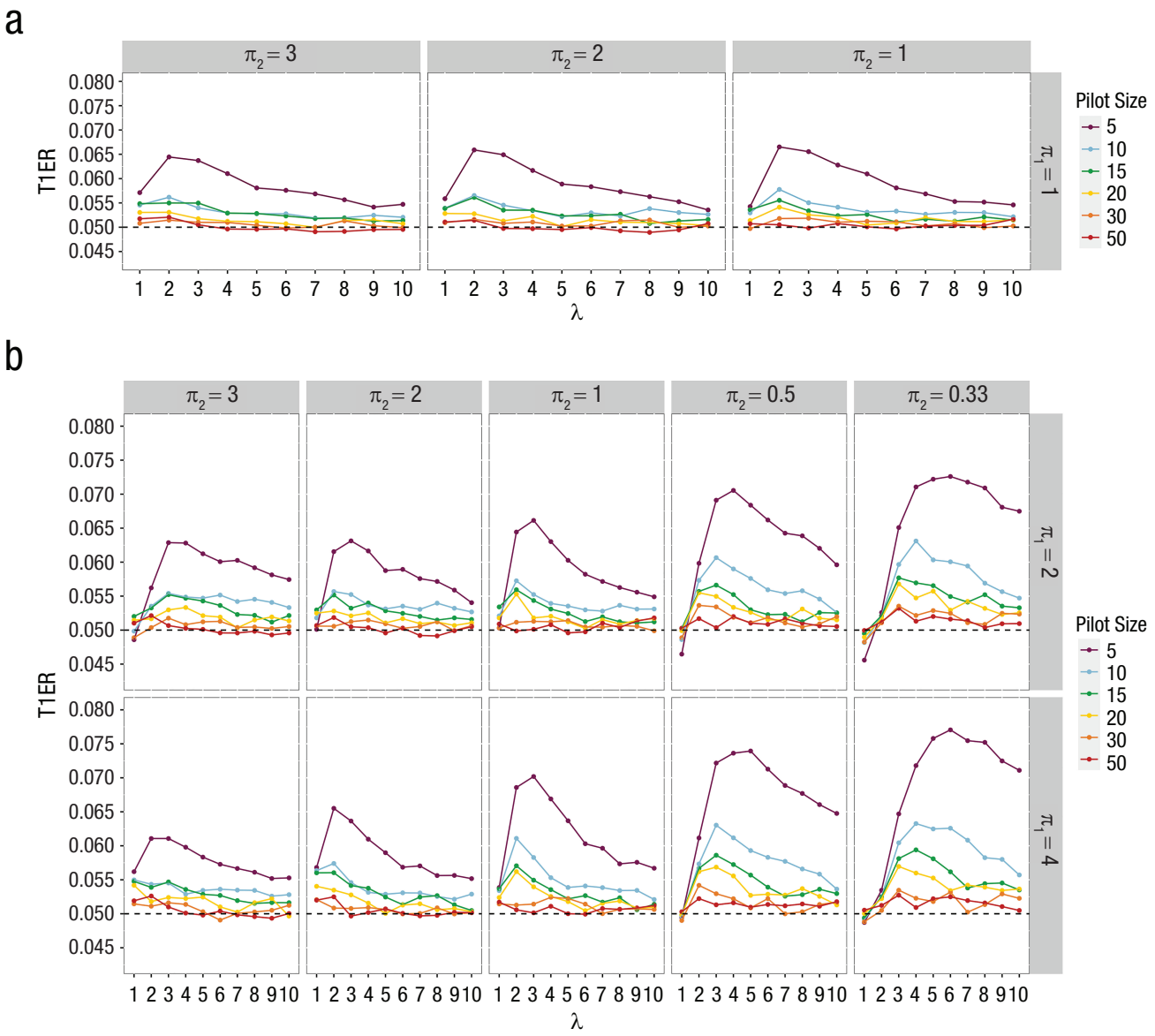

Fig. 4. The Type I error rate inflation of the internal-pilot-study design with unblinded variance estimation (a) for the Student *t* test when there is a common variance and (b) for the Welch *t* test when there are unequal variances. The inflation rates are displayed for different values of λ and internal-pilot size (equal for both groups) and investigated for different combinations of variance ratio (π_1) and sample-size ratio (π_2). The nominal α level of 5% is represented by the black dotted line.

smallest group has the largest variance (π_2 values smaller than 1), although this trend is reversed when the largest group is related to the largest variance (π_2 values larger than 1).

The T1ER levels are displayed for different values of λ and internal-pilot size (per group). A value of $\lambda = 1$ means that the total true required sample size (to uncover the SESOI with 80% power) is equally large as the already collected internal pilot. With $\lambda = 2$, the internal pilot includes half of the total true sample size. The larger λ is, the smaller the proportion of the total required sample size that is collected during the first stage is. The results show that for a fixed internal-pilot size, the T1ER reaches a peak when the multiplication factor λ is somewhere between 1 and 10. The average λ in which the T1ER is maximized is around 2.4. The reason why the T1ER returns to the nominal α level when λ is either very small or very large is that a design with a very large or very small internal pilot (respectively) relative to the total true required sample size resembles a fixed design without interim sample-size reestimation. In addition, it can be observed that the inflation decreases substantially when more data are collected during the internal-pilot stage. Using a large sample to obtain an interim variance estimate for sample-size reestimation leads to an actual T1ER close to the nominal α level.

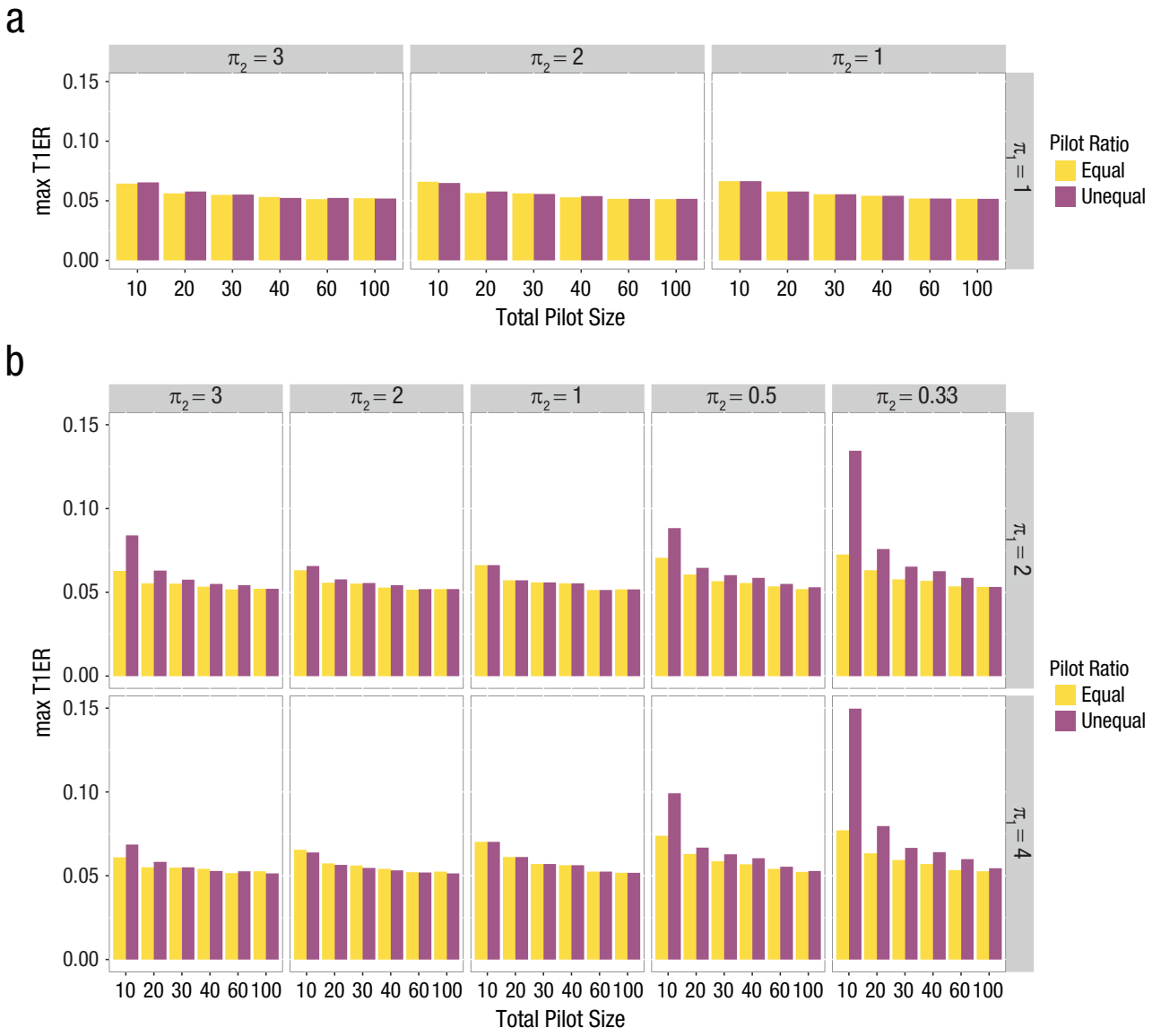

Fig. 5. Maximum Type I error rate for different combinations of π_1 and π_2 when the ratio of the internal pilot size of both groups is either 1 (i.e., equally sized internal pilots; yellow) or follows π_2 (i.e., unequally sized internal pilots when $\pi_2 \neq 1$; purple) (a) for the Student *t* test when there is a common variance and (b) for the Welch *t* test when there are unequal population variances.

Results: unequal internal-pilot size

The T1ER of the IPS procedure with the Student *t* test and the Welch *t* test was simulated for the same parameter configurations as in the simulation results described above, but now the first-stage data were collected with the same π_2 ratio used for the final sample sizes. Figure 5 compares the maximum T1ER when the ratio of the internal-pilot sizes is either always 1 (equal pilot sizes; i.e., the peaks of Fig. 4) or follows π_2 (unequal pilot sizes when $\pi_2 \neq 1$). The maximum Student T1ER remains the same whether the pooled variance is estimated based on equally sized or unequally sized internal-pilot groups (Fig. 5a). On the contrary, the maximum Welch T1ER is in most cases higher when estimating π_1 based on unequally sized pilot groups compared with equally sized internal pilots (Fig. 5b). This increase in inflation is due to the fact that for unequal pilots, the variance of one of the groups needs to be estimated with a smaller sample. For instance, suppose a total internal pilot of 20 observations and a π_2 of 0.33 or 3. The variance of each group can be estimated on each 10 observations when there are equal pilot sizes but can be based on only five observations for one group and 15 observations for the other group when the pilot ratio follows π_2. The increase in inflation is especially large when this small pilot of five observations also needs to estimate the largest variance (i.e., π_2 =0.33). When the size of the smallest

internal-pilot group increases (e.g., to 25 observations in case of total pilot of 100 and π_2. of 0.33 or 3), the additional inflation of having unequal pilot groups nearly disappears. Thus, there is no extra consequence on the Welch T1ER when collecting unequal internal-pilot groups when also the smallest of two pilot groups consists of a sufficiently large sample.

From simulation to practice

The above-discussed findings offer valuable insights into the behavior of the T1ER inflation of a study implementing the appropriate IPS procedure with unblinded variance estimation. The results demonstrate that when a researcher is willing/able to collect a large number of first-stage data in both groups to estimate the interim variance for sample-size reestimation, the T1ER inflation for both the Welch procedure and the Student procedure is negligible, regardless of how unequal (internal pilot) group sizes or population variances may be.[6] For example, collecting an internal pilot of at least 30 observations in both groups ensures that the T1ER will never be higher than 0.055. However, it is worth acknowledging that resource limitations, time constraints, or practical challenges may hinder the collection of a large internal-pilot sample for both groups. In cases in which one or both groups can provide only a (very) limited number of observations for unblinded interim variance estimation, the T1ER inflation may become noteworthy. We developed two R functions (one for the Student *t* test and one for the Welch *t* test) that can determine for each specific study with IPS design what the maximum T1ER inflation of the study is, accompanied by the corresponding adjusted α level to control the T1ER of the test at the originally planned nominal level α. This adjusted α level can be used as significance level in the final two-sample *t* test in studies in which a strict control of the α level is desired or necessary. When correction of the α level is deemed unnecessary, we believe it is both good practice and transparent to disclose the maximum T1ER inflation associated with the study in the respective publication.

R function

We provide two different R functions: the `find_adjalpha_welch()` function gives the maximum T1ER inflation and adjusted α level for a study performing the interim variance estimation with the π_1 estimator and using the Welch *t* test at the end and the `find_adjalpha_student()` function for the pooled variance estimator and the Student *t* test at the end. To be able to calculate the adjusted α level, the T1ER of the specific study applying the IPS design should first be identified. Remember from the simulation studies that the T1ER greatly depends on λ. The value of λ is never known in practice because λ relies on the true required sample size of the study (which is exactly what researchers try to estimate with the IPS design). The best one can do is search for the maximum T1ER for the specific study (i.e., searching for the peaks of the curves like in Fig. 4). In this way, regardless of the unknown λ, researchers can be informed about the worst-case T1ER inflation applicable to their study. This maximum inflation level can, in turn, be used to find the necessary adjusted α level. The adjusted α level can be determined using an iterative algorithm (Kieser & Friede, 2000), which is described in Appendix B. Because the adjusted α level departs from the maximized T1ER, using the adjusted α level as significance level in the final *t* test will lead to a conservative correction of the T1ER inflation.

The R function should be applied by the researcher at the interim analysis, after the first-stage data are collected. In this way, when deciding to use the adjusted α level for the final *t* test, the obtained adjusted α level can already be used as significance level in the formula for sample-size reestimation (which is beneficial for the power of the study). For the `find_adjalpha_welch()` function, the value for π_2, the size of the internal pilot that was collected in both groups, and the estimated variance ratio $\hat{\pi}_1$ based on the internal pilot needs to be specified at the interim analysis. When applying the `find_adjalpha_student()` function, π_1 is assumed to be 1. Only the value for π_2 and the internal-pilot size are needed to execute this function. For the set of parameters defined at the interim analysis, the two functions yield two essential outcomes: the maximized T1ER and the associated adjusted α level for subsequent application in the *t* test. The functions accomplish these objectives through a structured three-step process: (a) Use simulation to obtain the T1ER over a range of different λ values, (b) extract the maximized T1ER, and (c) use this maximized T1ER in the iterative algorithm to acquire the corresponding adjusted α level (Kieser & Friede, 2000). For a more detailed description of how the functions operate and the R code, see the Supplemental Material available online (also available at https://osf.io/mabrz/). To verify whether the functions can define the correct maximized T1ER (and thus also the correct adjusted α level) for a specific parameter set, the maximum T1ER found in the simulation study is compared with the maximized T1ER obtained with the functions (for both the Welch *t* test and the Student *t* test). For results, see Appendix C. Later in the article, we provide an example of implementing the R function in the IPS procedure.

IPS procedure for Student t *test or Welch* t *test?*

The choice of using the IPS procedure for the Student *t* test (with the pooled estimator) or the Welch *t* test (with the π_1 estimator) depends on the belief of the researcher

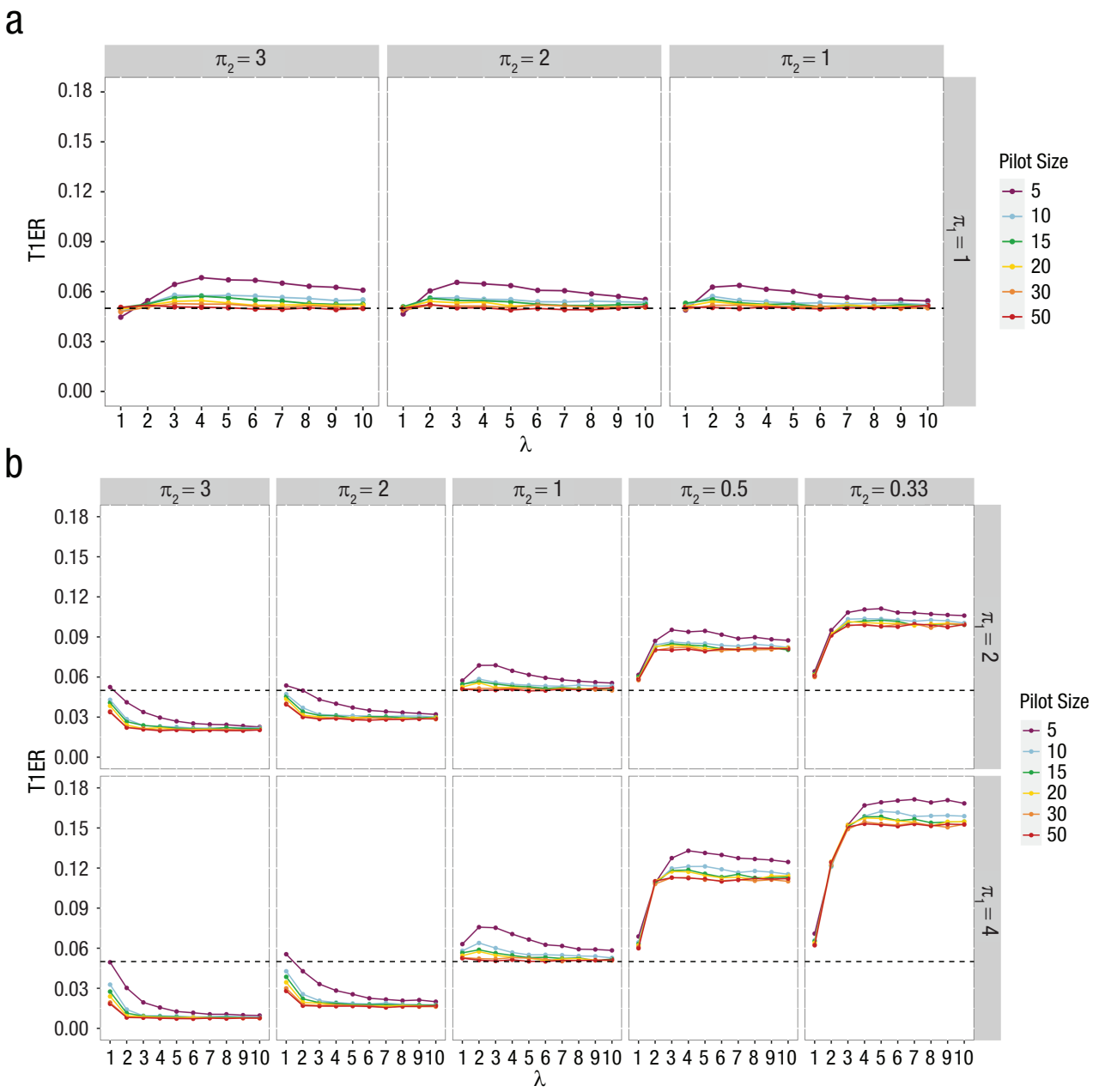

Fig. 6. The Type I error rate inflation of the internal-pilot-study design with unblinded variance estimation (a) for the Welch *t* test when there is a common variance and (b) for the Student *t* test when there are unequal variances. The inflation rates are displayed for different values of λ, internal-pilot size (equal for both groups), π_1, and π_2.

whether the homoscedasticity assumption for the study holds. However, researchers can never make this assumption with certainty; they can make only an informed guess. Figure 6 illustrates what happens to the T1ER of the two-sample *t* test when the "wrong" IPS procedure is used for the study, that is, the Welch procedure is used when, in reality, there are equal variances (Fig. 6a) or the Student procedure is used when, in reality, there are unequal variances (Fig. 6b). It can be observed that the Welch T1ER remains in the same (relatively small) range. The Student T1ER remains in the same range when the group sizes (and the internal-pilot group sizes) are equal in both groups but completely derails when group sizes are unequal. Then, the Student T1ER becomes either very conservative (for studies in which the largest group has the largest variance; $\pi_2 > 1$) or very liberal (for studies in which the smallest group has the largest variance; $\pi_2 < 1$). The further the variance ratio and group-size ratio move away from 1, the more the T1ER for the Student *t* test approaches zero when $\pi_2 > 1$ or exceeds the nominal α level when $\pi_2 < 1$. The conservative and liberal T1ER inflation rates of the Student *t* test found here

are in accordance with the findings of Delacre et al. (2017) for these scenarios within the fixed-sample design. One important takeaway from these results is that the IPS procedure with the pooled estimator and the Student *t* test can be "safely" used only in studies with equally sized final groups (and equally sized internal-pilot groups). Only then can it be assured that the IPS procedure for the Student *t* test can be employed without severe consequences on the T1ER and with an (more or less) accurate adjusted α level obtained with the `find_adjalpha_student()` function. When the researcher knows beforehand that large deviations from equal (internal) group sizes are expected, it is definitely safer to use the IPS procedure with the π_1 estimator and the Welch *t* test.

Power of the IPS Design

Method

To verify the property that the IPS design better guarantees the power of the study, a second simulation study was executed. The simulation code is available at https://osf.io/mabrz/. The power of the IPS procedure for both the Welch *t* test and Student *t* test was investigated for the same parameter configurations of group-size ratio (π_2), group-variance ratio (π_1), and internal-pilot size as used in the T1ER simulation. Because the power is not dependent on λ (as was the case for the T1ER), the power was for all scenarios simulated with a fixed population effect (i.e., difference in means) of 0.5. The same value of 0.5 was used as SESOI for the midcourse sample-size reestimation to be able to evaluate how well the IPS design reaches the intended power level. The population variance of Group 1 was for all scenarios set to 1 and was multiplied by π_1 for the variance value of Group 2. For each of the scenarios, the average empirical power and the range of the power over simulations (defined by a lower and an upper bound) were calculated. The simulation study was performed with a desired power of 80% and an α level that equals the adjusted α obtained with either the `find_adjalpha_student()` function (for the power of the Student IPS procedure) or the `find_adjalpha_welch()` function (for the power of the Welch IPS procedure) for the specific scenario. For more details on the simulation study, see Appendix D.

Results

Figure 7 displays the results for the average power (top) and the power range (bottom) for the Welch *t* test (Fig. 7a) and the Student *t* test (Fig. 7b). The average power of the Welch *t* test approaches the intended 80%,

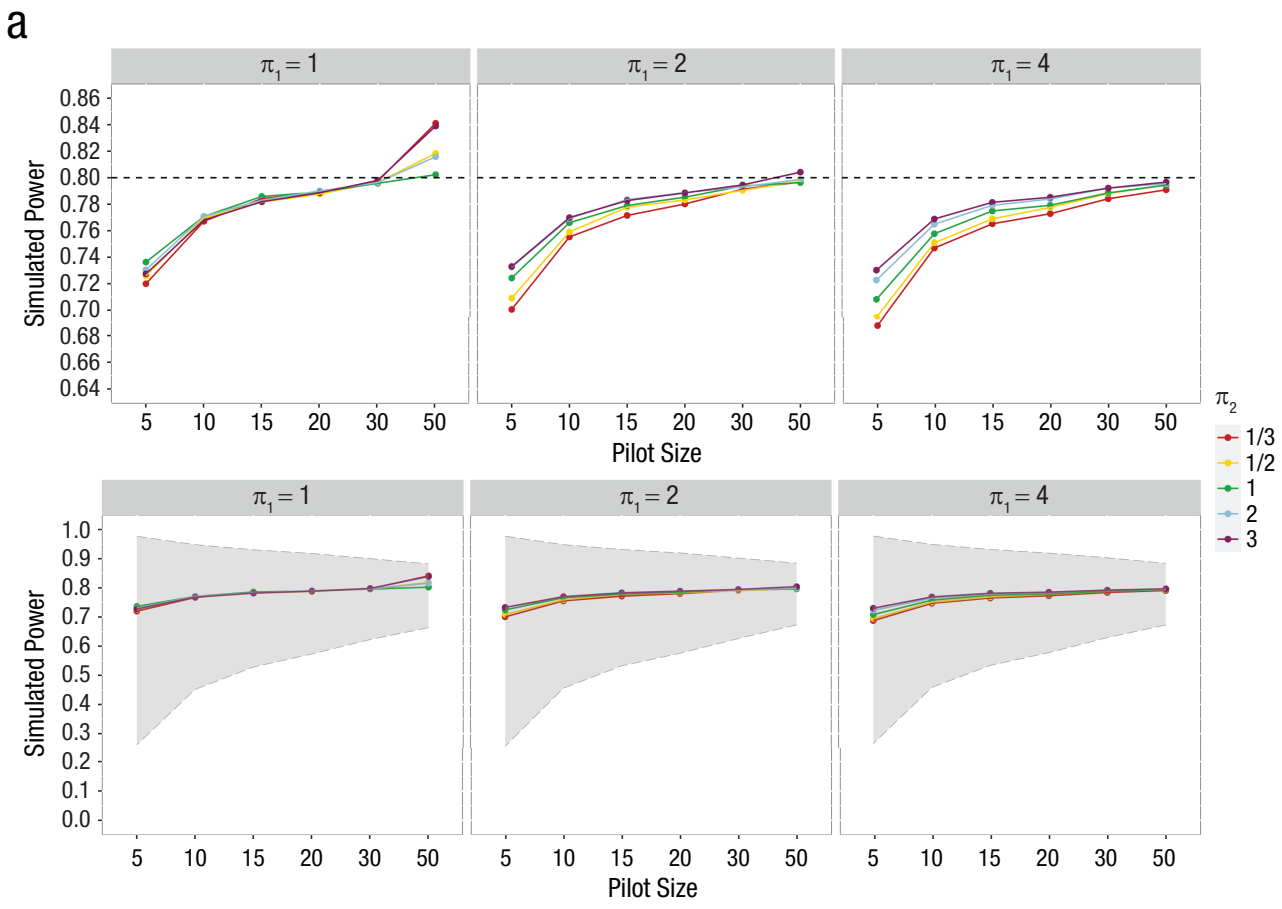

Fig. 7. *(continued on next page)*

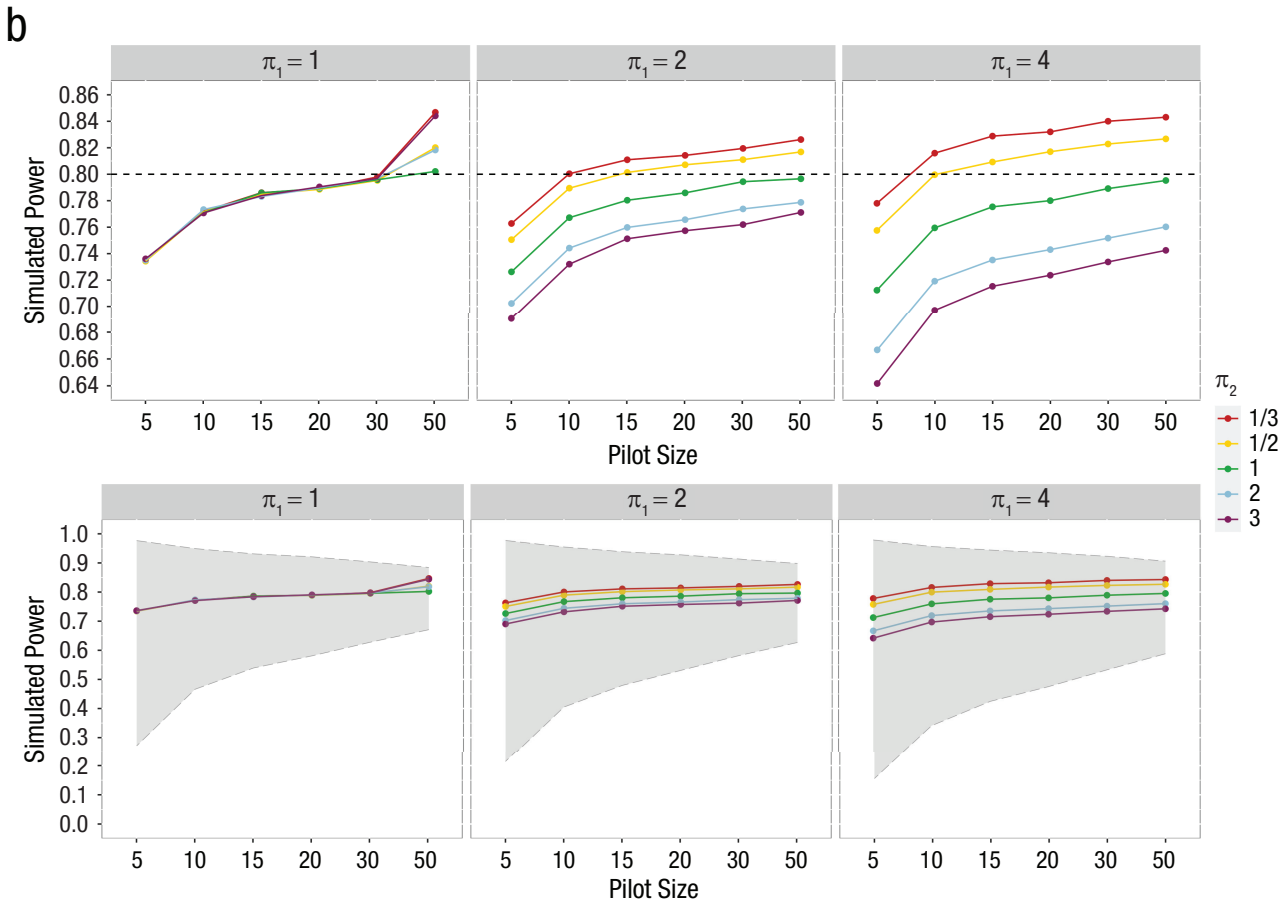

Fig. 7. The average simulated power with its minimum 10% lower bound and maximum 90% upper bound for (a) the Welch *t* test and (b) the Student *t* test for the same parameter configurations of π_1, π_2, and internal-pilot size (per group) as in the Type-1-error-rate simulation.

especially when internal-pilot size increases (e.g., around 70% for a small internal pilot of five observations per group, around 80% for 50 observations per group). There is a small influence of π_1 and π_2 on the resulting power: It is slightly smaller when the population variances of both groups are more unequal ($\pi_1 > 1$), lowest for unequally sized groups when the smallest group is related to the largest variance ($\pi_2 < 1$), and highest when the largest group is related to the largest variance ($\pi_2 > 1$). In some cases, the average empirical power rises substantially above the intended power level. This is because the collected internal pilot of one of the groups is larger than the true required sample size of that group. In the bottom row of Figure 7a, the 10% lower bound and 90% upper bound for the power are added. Because the difference in both the lower and the upper bounds is minimal for different values of π_2, only the minimum lower bound and maximum upper bound over all values of π_2 are plotted. In addition, over different values of π_1, the bounds remain very similar. The range of the power is large when the collected internal pilot is small and narrows considerably when the internal-pilot size increases. For example, the power of a single study performing sample-size reestimation after gathering five observations per group can be as low as 20% when the variance is underestimated or can be almost 100% when variance is overestimated. A severe under- or overestimation of the variance, and therefore of the required sample size, is less likely when more data are collected during the internal-pilot stage. An internal pilot of, for example, 30 observations per group will in almost all cases result in a study with a power not lower than 57% or higher than 92% to uncover the SESOI. The power of the Student *t* test is slightly better than the power of the Welch *t* test in the case of equal population variances ($\pi_1 = 1$; in terms of both average power and power range). When variances are unequal but group sizes remain equal ($\pi_2 = 1$), the power remains similar to the power of the Welch *t* test (we saw the same thing happening with the T1ER in this scenario). When both variances and group sizes are unequal, the power of the Student *t* test becomes distorted. A study in this scenario

```
> find_adjalpha_welch(n1.1 = 15, n1.2 = 15, est.pi1 = 1.67, pi2 = 0.5, alpha_nom = 0.05)
[1] "Refine sample size value for this parameter configuration"
  |++++++++++++++++++++++++++++++++++++++++++++++++++| 100% elapsed=33s
[1] "Searching for the adjusted alpha-level..."
  |++++++++++++++++++++++++++++++++++++++++++++++++++| 100% elapsed=30s
  |++++++++++++++++++++++++++++++++++++++++++++++++++| 100% elapsed=27s
$`Adjusted alpha`
[1] 0.04332597

$`Maximized T1ER`
[1] 0.05770212
```

Fig. 8. Example of the output of the `find_adjalpha_welch()` function.

performing the IPS procedure for the Student *t* test will be overpowered when the smallest group has the largest variance ($\pi_2 < 1$) and will be extra underpowered when the largest group has the largest variance ($\pi_2 > 1$).

Step-by-Step Tutorial With Practical Example

Below, we provide a comprehensive, step-by-step guide on the practical application of the IPS design with the purpose of correctly determining the final sample size of the study when treatment groups can be unblinded at the interim analysis and an adjusted α level is used to decide for significance. This tutorial is complemented by an illustrative example, which demonstrates the implementation of the IPS procedure employing the π_1 estimator and the Welch *t* test.

Step 1. The first step is optional and is mainly intended to gain insights into reasonable sizes for the internal pilot. Before the data collection starts, define an initial guess of the variance ratio π_1 and the population variance of Group 1 σ_1^2 (for the Welch *t* test) or the common variance σ^2 (for the Student *t* test). Plug this guess for π_1 and σ_1^2 into Equation 2 or the guess for σ^2 into Equation 1 to determine the initial sample-size estimate of Group 1 ($\hat{N}_1$). In case of unequal group sizes, the initial sample-size estimate of Group 2 ($\hat{N}_2$) can be obtained by multiplying initial $\hat{N}_1$ by π_2.

For example (for Welch), suppose a researcher is interested in uncovering a difference in means of 0.8 units with 80% power while testing at the 5% significance level and assumes that population variances are unequal in both groups. The researcher needs to set Group 1 to be twice as large as Group 2 ($\pi_2 = 0.5$). The variance of the outcome in Groups 1 and 2 are guessed to be 2 and 1.5, respectively, and hence the initial guess for π_1 is $1.5 / 2 = 0.75$. With Equation 2, the initial sample size of Group 1 is estimated to be 62. The estimated initial sample size of Group 2 is then $62 \times 0.5 = 31$.

Step 2. Decide on a feasible/reasonable amount of observations to collect for each group during the internal-pilot stage (optionally based on the initial sample-size calculation executed in Step 1). The internal-pilot size of both groups can be equal or unequal. More important is to try to collect a sufficient amount of first-stage data in both groups. After the first-stage data are collected, the variances of Group 1 and Group 2 are estimated separately and used to construct either the estimated variance ratio $\hat{\pi}_1$ (for the Welch *t* test) or the pooled variance estimate (for the Student *t* test).

For example (for Welch), based on the initial sample-size calculation, the researcher plans to estimate the variance in both groups after 15 observations per group have been collected. Suppose the variance of the outcome in this internal pilot turns out to be 1.5 (instead of 2) for Group 1 and 2.5 (instead of 1.5) for Group 2. The variance ratio $\hat{\pi}_1$ can be updated: $2.5 / 1.5 = 1.67$.

Step 3. Before performing the sample-size reestimation, obtain the adjusted α level necessary to correct for the T1ER inflation applicable for the study by executing the `find_adjalpha_welch()` function (for the Welch *t* test) or the `find_adjalpha_student()` function (for the Student *t* test).

For example (Welch), the `find_adjalpha_welch()` function is executed in R with an internal-pilot size for both groups of each 15 observations, an estimated variance ratio $\hat{\pi}_1$ of 1.67, a π_2 value of 0.5, and a nominal α level of 5%. The maximum T1ER inflation the researcher will have with this IPS design is 0.0577, which can be corrected by considering 0.0433 as adjusted α level (Fig. 8).

Step 4. Plug the estimated variance of Group 1 and the estimate $\hat{\pi}_1$ obtained on the first-stage data with the adjusted α level obtained with `find_adjalpha_welch()` into Equation 2 (for the Welch *t* test) or the estimated pooled variance based on the internal-pilot data with the adjusted α level obtained with `find_adjalpha_student()` into Equation 1 to acquire the reestimated

sample size of Group 1 ($\hat{N}_1$). The reestimated sample size of Group 2 is obtained by multiplying $\hat{N}_1$ with π_2.

For example (Welch), with a variance estimate of 1.5 for Group 1, an estimated variance ratio of 1.67, and an adjusted α level of .0433, the sample size of Group 1 is reestimated with Equation 2 to be 84. Consequently, the new estimated sample size of Group 2 is 42. In this scenario, there is an upward adjustment of the initially estimated sample size in both groups. However, it is also possible that the reestimated sample size is smaller than the amount of data already collected during the first stage (in either one or both of the groups). In that case, the final sample size of that group remains at the internal-pilot size.

Step 5. Collect the second-stage data for both groups, which is the remaining data on top of the internal-pilot data until the final (reestimated) sample size is reached. Afterward, the first- and second-stage data are pooled together, for both Group 1 and Group 2. The means of both groups are then compared with either the Welch *t* test or the Student *t* test.

For example (for Welch), for Group 1, an additional 69 observations need to be collected to reach the final reestimated sample size. For Group 2, this amount is 27. A total of 84 observations for Group 1 and 42 observations for Group 2 are then used in the final Welch *t* test. The *p* value obtained from this Welch *t* test should be compared with the adjusted significance level of 0.0433 to decide on significance.

Discussion

The IPS design that we proposed in this article is an adaptive design that can be used to determine the sample size of a study when the researcher is interested in uncovering an unstandardized population effect (of interest) but has no idea about the variance. In the IPS design, the nuisance parameters of a study are estimated on the first collected data and are then used to adapt (or reestimate) the final sample size of the study. In this article, we provide a general and complete overview of the IPS design for the independent-samples *t* test. More specifically, the procedure for performing the IPS design is explained for both studies assuming a common variance between the two groups (Student *t* test) and studies assuming unequal variances (Welch *t* test). For each of these two procedures, two different ways of estimating the interim variance are discussed: either blinded or unblinded. Blinded variance-estimation procedures might be especially useful when unblinding treatment groups comes with a lot of challenges (e.g., invoking an independent data monitoring committee for calculating the unblinded interim variance). However, IPS procedures with blinded variance estimation can be "safely" applied only when there are two equally sized (internal pilot) groups because blinded estimators can lead to very inaccurate estimates of the required sample size when variances between the two groups are in reality unequal. In addition, because blinded variance estimators potentially lead to unnecessarily large sample sizes and unblinding treatment groups is often not as complex in psychological research compared with the stricter clincial-trials context, the psychological researcher might estimate the interim variance unblinded to obtain an unbiased estimate of the true required sample size. Unfortunately, unblinding comes with a small, sometimes nonnegligible inflation of the T1ER. Next to quantifying this inflation for both the Student *t* test and the Welch *t* test, in this article, we offer R functions that provide an adjusted α level to correct for the maximum inflation.

Switching from the traditional fixed-sample design with a priori sample-size calculation to this adaptive design with midcourse sample-size calculation has several benefits. First and foremost, the IPS design offers the possibility to greatly improve the power of a study. The significant uncertainty regarding the value of the population variance(s) that is present at the start of the study when performing an a priori sample-size calculation can be solved with this design. Although an IPS design with a very small internal pilot of, for example, five or 10 observations per group does not have a real benefit on the power over an a priori sample-size calculation with a small external-pilot sample (except that data can be retained), the true advantage of the IPS design is that precise variance estimates can be obtained without wasting any data, leading to properly powered studies. The IPS design should therefore be considered over the traditional a priori power-analysis approach when the study has the possibility to collect a large(r) amount of internal-pilot data. For example, a study collecting an internal pilot of 15 observations per group will have, on average, a power around 77% and will unlikely have a power lower than 50% (when the desired power level is 80%), whereas the power with the a priori power-analysis approach can easily drop to 29% when the variance estimate considered in the sample-size calculation is only 2 times smaller than the true variance. When the interim sample-size calculation is performed after a very small amount of internal-pilot data, also the IPS design cannot solve the problem of a possible underpowered study because small samples will always lead to inaccurate variance estimates and therefore to unsatisfactory power. An alternative adaptive design, in which the variance parameter is continuously monitored during the course of the study and the data collection is stopped only when the estimate of the variance parameter fulfills a stopping criterion (e.g., a bound on the standard error of the mean difference), might be an even better alternative to ensure decent power for a study (Friede & Miller,

2012; Mehta & Tsiatis, 2001). Another somewhat similar alternative is the sequential probability ratio test, in which the data are sequentially analyzed by continuously calculating the likelihood ratio, indicating how likely the observed data occur under a specific alternative hypothesis versus the null hypothesis until enough evidence is found in one of both directions (Schnuerch & Erdfelder, 2020). The disadvantage of both of these approaches is that it is impossible to know beforehand how large the final sample size of the study will be. When this practical difficulty is not of concern, these approaches are definitely worth considering.

A second benefit of the IPS design is that the researcher should not be too worried about the T1ER inflation of the final Welch *t* test or Student *t* test. As shown in the T1ER simulation studies, the inflation is very small when the internal pilot for both groups is large but can still be corrected for with the developed R functions `find_adjalpha_welch()` and `find_adjalpha_student()` when desired. Finally, the process of determining the sample size of a study is easier with this flexible design compared with the traditional a priori power analysis. Before the study starts, researchers need to specify only the minimal difference in means of interest, and they should not be concerned about an accurate a priori guess of the unknown variance. Even though the burden of correctly specifying the variance a priori disappears, defining an unstandardized SESOI remains a difficult task, particularly in the absence of prior research or established theories on the topic. A first step forward could be to discuss with peers which effect sizes can be considered meaningful. Even though formulating a SESOI is not straightforward, it is key to reflect on what effect sizes matter. Studies powered for an unstandardized SESOI control the Type II error rate for effect sizes that matter (when assuming the variance is correctly estimated), which is not the case when the study is powered for an effect size estimated on previous data or pilot data.

We recognize that in psychological research, it is less common to separately define a specific mean difference of interest and variance estimate for sample-size calculations because measurement scales can differ between studies. However, we emphasize that this approach offers clear insights into the study's statistical power. Consider a scenario in which a standardized effect of 0.8 (Cohen's *d*) is used to determine the required sample size of the study. When the variance in the data turns out to be very large, the power of the study may be limited to detecting only substantial mean differences—possibly exceeding the threshold of practical significance. In such instances, the study's utility may be compromised. Hence, a precise and accurate estimation of the variance is essential for conducting efficient and well-powered research, which can be achieved with the IPS design. Researchers who prefer to conceptualize their studies in terms of standardized effects and still aim to power their research for a specific standardized effect size can still benefit from the IPS design. They can use the first couple of collected data to estimate the variance, deduct the raw mean difference needed to achieve the desired standardized effect, and perform the midcourse sample-size calculation based on these tailored values. Without excluding the use of standardized effect sizes, the method described in this article offers a more direct way to account for the variable's distribution because it disentangles the raw effect and the variance.

In this article, we explained and investigated the IPS design only for the two-sample *t* test, but this design can also easily be implemented in studies relying on other types of statistical analyses. The sample-size calculation for the specific analysis just needs to be performed midcourse (instead of a priori), where the nuisance parameters for the sample-size calculation are estimated on the first-stage data. Whereas previous (and this) research investigated only the T1ER inflation for the IPS design in the case of the two-sample *t* test, we conject that the T1ER inflation for other types of statistical analyses will remain in the same relatively small range.

Appendix A

Detailed methods of Type I error rate simulation

In the Type I error rate (T1ER) simulation study, the different steps of the internal-pilot-study (IPS) design as described in Figure 2 are mimicked on data simulated under the null hypothesis of no difference in means: (1) Collect for each group an internal pilot of a certain size, (2) use these internal-pilot data to estimate the variance of both groups and construct either the pooled variance (Student *t* test) or $\hat{\pi}_1$ (Welch *t* test), (3) calculate the final sample size per group by plugging in the estimated pooled variance in Equation 1 or $\hat{\pi}_1$ in Equation 2, (4) determine per group how much data need to be collected on top of the internal pilot to reach the reestimated final sample size and collect these data, and (5) at the end of data collection, perform the correct two-sample *t* test on the simulated data (Student or Welch). This process was repeated 100,000 times, and the empirical T1ER was calculated by counting how often the null hypothesis was incorrectly rejected (p value $< \alpha$).

To investigate the T1ER within the different scenarios, the simulation study was executed for different combinations of π_1 and π_2 (Fig. 9). In the equally sized groups scenario ($\pi_2 = 1$), the data of both groups were simulated with a variance ratio of 1 (equal variances), 2, or 4 (unequal variances). In the unequally sized groups

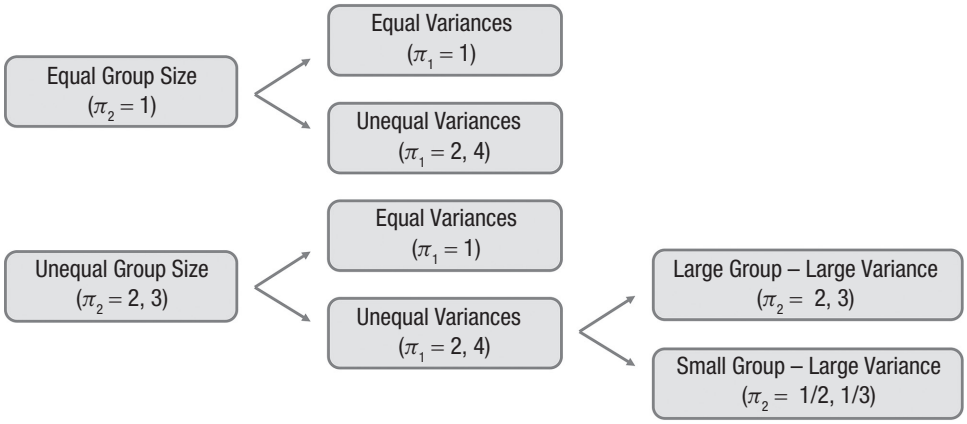

Fig. 9. The different π_1 and π_2 combinations investigated in the simulation.

scenario ($\pi_2 \neq 1$), the same variance ratios were used. However, in the case of unequal variances, a distinction was made between the largest group having the largest variance ($\pi_2 = 2, 3$) and the smallest group having the largest variance ($\pi_2 = .50, .33$). Within each combination of π_1 and π_2, multiple sizes of the internal pilot were considered. The total internal-pilot size (for both groups together) consisted of either 10, 20, 30, 40, 60, or 100 observations. In the first simulation study, the internal pilot size of both groups was always equal, regardless of the value of π_2 (resulting in an internal pilot of either 5, 10, 15, 20, 30, or 50 observations per group). In the second simulation study, the same π_2 ratio present in the final sample size was used to divide the total internal-pilot size across both groups. For each configuration of π_1, π_2 and total internal pilot size, λ was varied from 1 to 10. λ represents the ratio of the true required sample size (to detect the SESOI used in the midcourse sample-size calculation in Step 3) relative to the size of the internal pilot. The value of the SESOI was chosen in such a way that, in combination with the variance with which the data of both groups were simulated, the resulting true required sample size was larger than the collected internal pilot by these different values of λ. Finally, the α level and the desired power were fixed at 5% and 80%, respectively.

Appendix B

Iterative algorithm of Kieser and Friede (2000)

Because the maximized T1ER (α_{act}^{max}) is a monotone function of α (for fixed values of the power and internal pilot size), the adjusted α level (α_{adj}) can be found by solving the following equation:

$$\alpha_{act}^{max}\left(\alpha_{adj}\right) = \alpha. \quad (3)$$

To solve Equation 3, an iterative algorithm can be used in which the maximized actual T1ER is identified for each newly obtained adjusted α level. The final adjusted α level is that value for which the corresponding maximized actual T1ER does not exceed the originally planned nominal level. To clarify the different steps of the algorithm, it is applied to a concrete example in the user manual of the R functions (see the Supplemental Material available online). In general, the iterative algorithm can be summarized as follows:

1. Determine the initial value for α_{adj} by first calculating α_{act}^{max} for the nominal α level and then filling in following formula: $\alpha_{adj}^{(0)} = \alpha \times \frac{\alpha}{\alpha_{act}^{max}(\alpha)}$.
2. In the next k steps of the algorithm ($k \geq 1$), the new adjusted α level is obtained by following formula: $\alpha_{adj}^{(k)} = \alpha_{adj}^{(k-1)} \times \frac{\alpha}{\alpha_{act}^{max}(\alpha_{adj}^{(k-1)})}$, where the α_{act}^{max} is every time identified for the previous adjusted α level.
3. The algorithm stops when the difference between the maximized actual T1ER of the latest adjusted α level ($\alpha_{act}^{max}(\alpha_{adj}^{(s)})$) and the nominal level α is less than a prespecified error term ϵ. The final adjusted α level is then defined by $\alpha_{adj}^{(s)}$.

Appendix C

Validation `find_adjalpha()` *functions*

To assess the performance of the `find_adjalpha_welch()` function (for the IPS procedure with the Welch *t* test) and the `find_adjalpha_student()` function (for the IPS procedure with the Student *t* test), the maximum T1ER found in the simulation study is compared with the maximum T1ER obtained from the

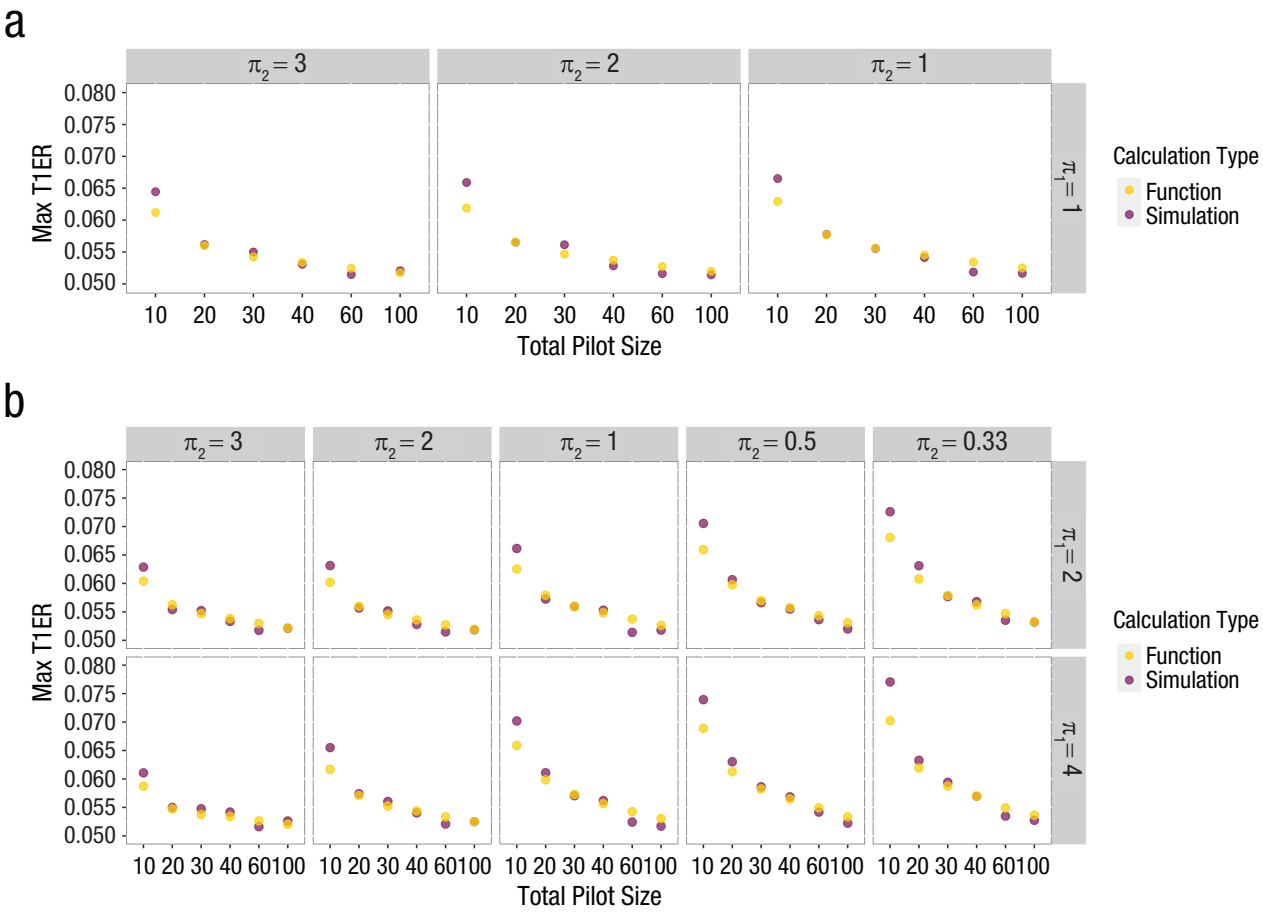

Fig. 10. The maximized Type I error rate (T1ER) inflation obtained in the T1ER simulation study (purple) and with the R function `find_adjalpha()` (yellow) (a) for the Student *t* test and (b) for the Welch *t* test.

function. For this purpose, the `find_adjalpha()` functions were used to compute the maximized inflation levels for the same configurations of π_1, π_2 and internal pilot size as used in the T1ER simulation. The function was executed with a precision of 1,000,000 simulations and for a nominal α level of .05. Figure 10 displays the maximized inflation values obtained from both the simulation study and the R function for the different parameter configurations, for both the Student *t* test (Fig. 10a) and the Welch *t* test (Fig. 10b). For a very small internal-pilot size (i.e., five observations per group), the function slightly underestimates the maximum T1ER. However, for all other pilot sizes (and for all combinations of π_1 and π_2), the functions successfully estimate the true maximized inflation (as obtained from the simulation study).

Appendix D

Detailed methods of power simulation

The same steps as in the T1ER simulation to mimick the IPS design with unblinded variance estimation were followed to demonstrate the power of the Welch *t* test and the Student *t* test. The only difference with the procedure of the T1ER simulation study is that (a) the data were generated under the alternative hypothesis instead of the null hypothesis and (b) the adjusted α is used at the interim analysis to perform the sample-size reestimation. The power was explored for a combination of the same parameter configurations of π_1, π_2 (see Fig. 9), and internal-pilot sizes (i.e., five, 10, 15, 20, 30, and 50 per group) as in the T1ER simulation. Instead of investigating the power within each of these parameter configurations for different values of λ (as was done in the T1ER simulation), the effect size with which the data were simulated and the SESOI used in the sample-size recalculation were for all parameter configurations fixed at a value of 0.5. The populations variance of Group 1 was set to the value of 1 and was multiplied by π_1 to obtain the population variance of Group 2. The SESOI used for sample-size recalculation was set to the same value as the true effect size with which the data were simulated to investigate how well the actual power approximates the predefined power level. Power will be even higher when the true effect is larger than the SESOI and will be smaller for the opposite situation.

The average empirical power was retrieved by calculating how often the null hypothesis was correctly rejected by comparing the obtained p value with the adjusted α level for that parameter configuration (p value < α_{adj}). The range of the power over simulations was defined by calculating a 10% lower power bound and 90% upper power bound. These bounds were obtained by determining the 10% lower and 90% upper quantile of the sampling distribution of the sample variance and using these variance estimates during the sample-size recalculation of the simulation. The nominal significance level was set to 5%, the intended power to 80%, and the number of simulations to 100,000.

Transparency

Action Editor: Yasemin Kisbu-Sakarya
Editor: David A. Sbarra
Author Contribution(s)

Lara Vankelecom: Conceptualization; Formal analysis; Investigation; Methodology; Software; Visualization; Writing – original draft; Writing – review & editing.

Tom Loeys: Conceptualization; Investigation; Methodology; Software; Supervision; Validation; Writing – review & editing.

Beatrijs Moerkerke: Conceptualization; Investigation; Methodology; Software; Supervision; Validation; Writing – review & editing.

Declaration of Conflicting Interests

The author(s) declare that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

T. Loeys and B. Moerkerke acknowledge the Research Foundation Flanders for financial support (Grant G023623N).

Open Practices

This article has received the badge for Open Materials. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

ORCID iDs

Lara Vankelecom https://orcid.org/0000-0003-3213-9528
Tom Loeys https://orcid.org/0000-0003-4551-5502
Beatrijs Moerkerke https://orcid.org/0000-0003-1580-547X

Notes

1. Tutorials on sample-size calculation for other types of statistical analyses also exist (e.g., for analysis of variance, see Lakens & Caldwell, 2021; for multilevel regression models in longitudinal studies, see Lafit et al., 2021; for the independent two-group studies, see Kovacs et al., 2022; or for less common statistical tests and techniques with simulation, see Arnold et al., 2011).
2. The Student t test uses following test statistic: $t = (\bar{x}_1 - \bar{x}_2)/\sqrt{s^2_{pooled} \times (1/n_1 + 1/n_2)}$, where $\bar{x}_1$ and $\bar{x}_2$ represent the sample means of the two groups and the pooled variance is the estimator of the true common variance.
3. Note that by using the standard normal instead of the t distribution, Equations 1 and 2 give an asymptotic approximation of the required sample size.
4. The Welch t test calculates following test statistic: $t = (\bar{x}_1 - \bar{x}_2)/\sqrt{s_1^2/n_1 + s_2^2/n_2}$, with $\bar{x}_1$ and $\bar{x}_2$ the two sample means and s_1^2 and s_2^2 the separate estimates of the two unequal population variances.
5. There are also designs that allow to unblind the effect at the interim analysis and use it to perform an interim hypothesis test to stop the study early for efficacy or futility (Wassmer & Brannath, 2016), but this falls out of the scope of this article.
6. The simulation study verified only the T1ER for scenarios in which the final sample size in one group is no more than 3 times larger than in the other group and the population variance in one group is no more than 4 times larger than in the other group.

References

Aberson, C. L. (2019). *Applied power analysis for the behavioral sciences*. Routledge.

Albers, C., & Lakens, D. (2018). When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias. *Journal of Experimental Social Psychology*, *74*, 187–195.

Arnold, B. F., Hogan, D. R., Colford, J. M., & Hubbard, A. E. (2011). Simulation methods to estimate design power: An overview for applied research. *BMC Medical Research Methodology*, *11*, Article 94. https://doi.org/10.1186/1471-2288-11-94

Bakker, M., Hartgerink, C. H., Wicherts, J. M., & van der Maas, H. L. (2016). Researchers' intuitions about power in psychological research. *Psychological Science*, *27*(8), 1069–1077.

Bakker, M., Van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*(6), 543–554.

Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods*, *43*(3), 666–678.

Birkett, M. A., & Day, S. J. (1994). Internal pilot studies for estimating sample size. *Statistics in Medicine*, *13*(23–24), 2455–2463.

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition*, *2*(1), Article 16. https://doi.org/10.5334/joc.72

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Erlbaum.

Cohen, J. (1990, August). *Things I have learned (so far)* [Paper presentation]. 98th Annual convention of the American Psychological Association, Boston, MA.

Cummings, P. (2011). Arguments for and against standardized mean differences (effect sizes). *Archives of Pediatrics & Adolescent Medicine*, *165*(7), 592–596.

Delacre, M., Lakens, D., & Leys, C. (2017). Why psychologists should by default use Welch's t-test instead of student's *t*-test. *International Review of Social Psychology*, *30*(1), Article 92-101. https://doi.org/10.5334/irsp.82

Ekstrøm, C. T. (2020). *Mess: Miscellaneous esoteric statistical scripts* (R package version 0.5.7). https://CRAN.R-project.org/package=MESS

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLOS ONE*, *4*(5), Article e5738. https://doi.org/10.1371/journal.pone.0005738

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*(3), 891–904.

Fraley, R. C., & Vazire, S. (2014). The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PLOS ONE*, *9*(10), Article e109019. https://doi.org/10.1371/journal.pone.0109019

Friede, T., & Kieser, M. (2001). A comparison of methods for adaptive sample size adjustment. *Statistics in Medicine*, *20*(24), 3861–3873.

Friede, T., & Kieser, M. (2006). Sample size recalculation in internal pilot study designs: A review. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, *48*(4), 537–555.

Friede, T., & Miller, F. (2012). Blinded continuous monitoring of nuisance parameters in clinical trials. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *61*(4), 601–618.

Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General*, *141*(1), 2–18.

Gould, A., & Shih, W. J. (1992). Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance. *Communications in Statistics: Theory and Methods*, *21*(10), 2833–2853.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLOS Medicine*, *2*(8), Article e124. https://doi.org/10.1371/journal.pmed.0020124

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532.

Kieser, M., & Friede, T. (2000). Re-calculating the sample size in internal pilot study designs with control of the Type I error rate. *Statistics in Medicine*, *19*(7), 901–911.

Kieser, M., & Friede, T. (2003). Simple procedures for blinded sample size adjustment that do not affect the Type I error rate. *Statistics in Medicine*, *22*(23), 3571–3581.

Koehler, K. J. (1983). A simple approximation for the percentiles of the *t* distribution. *Technometrics*, *25*(1), 103–105.

Kovacs, M., van Ravenzwaaij, D., Hoekstra, R., & Aczel, B. (2022). SampleSizePlanner: A tool to estimate and justify sample size for two-group studies. *Advances in Methods and Practices in Psychological Science*, *5*(1). https://doi.org/10.1177/25152459211054059

Lafit, G., Adolf, J. K., Dejonckheere, E., Myin-Germeys, I., Viechtbauer, W., & Ceulemans, E. (2021). Selection of the number of participants in intensive longitudinal studies: A user-friendly shiny app and tutorial for performing power analysis in multilevel regression models that account for temporal dependencies. *Advances in Methods and Practices in Psychological Science*, *4*(1). https://doi.org/10.1177/2515245920978738

Lakens, D. (2022). Sample size justification. *Collabra: Psychology*, *8*(1), Article 33267. https://doi.org/10.1525/collabra.33267

Lakens, D., & Caldwell, A. R. (2021). Simulation-based power analysis for factorial analysis of variance designs. *Advances in Methods and Practices in Psychological Science*, *4*(1). https://doi.org/10.1177/2515245920951503

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269.

Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *The American Statistician*, *55*(3), 187–193.

Mehta, C. R., & Tsiatis, A. A. (2001). Flexible sample size considerations using information-based interim monitoring. *Drug Information Journal*, *35*, 1095–1112.

Miller, F. (2005). Variance estimation in clinical studies with interim sample size reestimation. *Biometrics*, *61*(2), 355–361.

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference: Part I. *Biometrika*, *20A*(1/2), 175–240.

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, *7*(6), 531–536.

Polit, D. F., & Beck, C. T. (2004). *Nursing research: Principles and methods*. Lippincott Williams & Wilkins.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

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

Stein, C. (1945). A two-sample test for a linear hypothesis whose power is independent of the variance. *The Annals of Mathematical Statistics*, *16*(3), 243–258.

Wacholder, S., Chanock, S., Garcia-Closas, M., El Ghormli, L., & Rothman, N. (2004). Assessing the probability that a positive report is false: An approach for molecular epidemiology studies. *Journal of the National Cancer Institute*, *96*(6), 434–442.

Wassmer, G., & Brannath, W. (2016). *Group sequential and confirmatory adaptive designs in clinical trials* (Vol. 301). Springer.

Wittes, J., & Brittain, E. (1990). The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine*, *9*(1–2), 65–72.

Wittes, J., Schabenberger, O., Zucker, D., Brittain, E., & Proschan, M. (1999). Internal pilot studies I: Type I error rate of the naive t-test. *Statistics in Medicine*, *18*(24), 3481–3491.

Zucker, D. M., Wittes, J. T., Schabenberger, O., & Brittain, E. (1999). Internal pilot studies II: Comparison of various procedures. *Statistics in Medicine*, *18*(24), 3493–3509.