Journal of Experimental Social Psychology 74 (2018) 187–195

Contents lists available at ScienceDirect

Journal of Experimental Social Psychology

journal homepage: www.elsevier.com/locate/jesp

When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias☆

Casper Albers[a,*,1,2], Daniël Lakens[b,2,3]

[a] University of Groningen, The Netherlands
[b] Eindhoven University, The Netherlands

ARTICLE INFO

Keywords:
Effect size
Power analysis
Follow-up bias
Eta-squared
Omega-squared
Epsilon-squared

ABSTRACT

When designing a study, the planned sample size is often based on power analyses. One way to choose an effect size for power analyses is by relying on pilot data. A-priori power analyses are only accurate when the effect size estimate is accurate. In this paper we highlight two sources of bias when performing a-priori power analyses for between-subject designs based on pilot data. First, we examine how the choice of the effect size index (η^2, ω^2 and ε^2) affects the sample size and power of the main study. Based on our observations, we recommend against the use of η^2 in a-priori power analyses. Second, we examine how the maximum sample size researchers are willing to collect in a main study (e.g. due to time or financial constraints) leads to overestimated effect size estimates in the studies that are performed. Determining the required sample size exclusively based on the effect size estimates from pilot data, and following up on pilot studies only when the sample size estimate for the main study is considered feasible, creates what we term *follow-up bias*. We explain how follow-up bias leads to underpowered main studies.

Our simulations show that designing main studies based on effect sizes estimated from small pilot studies does not yield desired levels of power due to accuracy bias and follow-up bias, even when publication bias is not an issue. We urge researchers to consider alternative approaches to determining the sample size of their studies, and discuss several options.

1. Introduction

It is common practice in psychological and behavioral research to express the results of a quantitative study in at least two numbers: One expressing the probability or likelihood of data under specified statistical models, usually through a *p*-value or Bayes factor, and one expressing the magnitude of the effect, often through a (standardized) effect size (ES). Reporting effect size estimates serves various purposes, one of which is facilitating cumulative science by allowing other researchers to use the effect size estimate in a-priori power analyses (Cohen, 1988). Power analyses can be used to design studies that have a desired probability of observing a statistically significant effect, assuming there is a true effect of a specified size. However, a-priori power analyses are only accurate when the effect size estimate is accurate. It has been pointed out that effect sizes reported in the literature are known to be inflated due to publication bias, and this widespread bias in reported effect sizes is a challenge when performing a-priori power analyses based on published research.

In this manuscript, we focus on two other sources of bias in power analyses that play an important role in power analysis even when publication bias and researchers' degrees of freedom do not influence effect size estimates (e.g., when researchers perform their own pilot study). These sources of bias point out clear limitations of the common practice to use the effect size from a pilot study to determine the sample size of a follow-up study through an a-priori power analysis. First, we will discuss the relatively straightforward matter of the impact of a biased effect size estimator (η^2), compared to less biased effect size estimators (ε^2 and ω^2) on the sample size estimate in power analyses. Second, we examine a source of bias which we refer to as *follow-up bias*. Effect size estimates vary around the true effect size. Even without publication bias, researchers are more likely to follow-up on initial studies that yielded higher effect size estimates than initial studies that

☆ The Supplementary Material, including the full R code for the simulations and plots can be obtained from the Open Science Framework https://osf.io/zq9mg/.

* Corresponding author: Grote Kruisstraat 2/1, 9712, TS, Groningen, The Netherlands.
E-mail address: c.j.albers@rug.nl (C. Albers).
[1] Department of Psychology, University of Groningen, The Netherlands.
[2] Both authors contributed equally to this manuscript.
[3] School of Innovation Sciences, Eindhoven University of Technology, The Netherlands.

http://dx.doi.org/10.1016/j.jesp.2017.09.004
Received 18 April 2016; Received in revised form 1 September 2017; Accepted 10 September 2017
Available online 16 October 2017

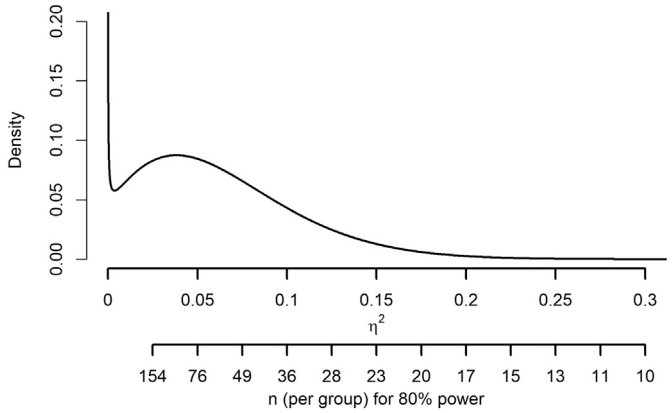

Fig. 1. Distribution of η^2 for a between-subjects *t*-test with a sample size (per group) of $n = 50$, and a medium true population effect size ($\eta^2 = 0.0588$). The lower x-axis indicates the *n* per group required to achieve .80 power based on the observed effect size indicated by the upper x-axis. Reprinted from http://dx.doi.org/10.6084/m9.figshare.4877414, CC-BY4.0.

yielded lower effect size estimates (cf. Greenwald, 1975), simply because these studies require less resources to observe a statistically significant result in the expected direction. We examine how this understandable behavior leads to an overestimation of the true effect size, on average, when performing a-priori power analyses, and thus leads to follow-up studies that are underpowered. Based on these observations, we argue against recent recommendations (Sakaluk, 2016) to use small pilot studies to explore effects. In the discussion, we offer some general recommendations to design well-powered studies.

2. Eta-squared, Epsilon-squared, and Omega-squared

In experimental psychology, it is extremely common to perform studies where participants are randomly assigned to different conditions, and analyze the results using analysis of variance (ANOVA) or (unpaired) *t*-tests (where a *t*-test is mathematically identical to a one-way ANOVA with two groups). We will illustrate our main points using ANOVA and the related effect sizes, but our conclusions generalize to other effect sizes and statistical tests. In one-way ANOVA, all analyses are based on the following decomposition of the variance. The total variance of all measurements together, ${\sigma_T}^2$, is split into a part that can be attributed to group-membership (${\sigma_B}^2$) and a part that can not (${\sigma_W}^2$):

$$\sigma_T^2 = \sigma_B^2 + \sigma_W^2.$$

The subscripts W, B, and T indicate 'within' samples, 'between' samples, and 'total'. Equivalently, one can decompose the so-called sums of squares:

$$SS_T = SS_B + SS_W.$$

One of the most common effect size indices in one-way ANOVA is eta-squared (η^2), which describes the proportion of variance that is explained by group membership. It dates back to {Citation}Pearson (1911), who introduced it in a regression context, and to Fisher (1928), who used it in the ANOVA context. In statistical packages such as SPSS, eta-squared is the default effect size measure. Eta-squared is an upwardly biased estimate of the true population effect size, and two alternative effect size indices have been suggested that are less biased, namely epsilon-squared (ε^2, Kelley, 1935) and omega-squared (ω^2, Hays, 1963). For background reading on these (and other) indices, we refer to Levine and Hullett (2002), Okada (2013) and McGrath and Meyer (2006), and the references therein. Eta-squared, epsilon-squared, and omega-squared are defined as follows[4]:

$$\eta^2 = \frac{\sigma_B^2}{\sigma_T^2} = \frac{SS_B}{SS_T},$$

$$\varepsilon^2 = \frac{SS_B - df_B \times MS_W}{SS_T},$$

$$\omega^2 = \frac{SS_B - df_B \times MS_W}{SS_T + MS_W},$$

where, using standard ANOVA-notation, *SS*, *MS* and *df* denote the sum-of-squares, mean sum-of-squares, and degrees of freedom. From these effect size estimates, the well-known Cohen's *d* and Cohen's *f* can be estimated (Cohen, 1988). For population effect sizes Cohen (1988, p. 276) states that $d = 2f$ with $f^2 = \eta^2/(1 - \eta^2)$. An unbiased estimate of Cohen's *d* is called Hedges' *g* (see Lakens, 2013), and recommendations in this article concerning the use of ω^2 and ε^2 instead of η^2 extend to the use of Hedges' *g* instead of Cohen's *d*.

Alternative formulas for these effect sizes, where the computation is based only on the *F*-value and the degrees of freedom, are given in Appendix A.

These indices are estimators of the unknown true population effect size and, as such, contain possible bias and variability. It is well-known that η^2 has more bias than the other two indices, but the other two indices have more variability (cf. Albers, 2015; Lakens, 2015; Okada, 2013). The amount of bias and variability of these indices depends on the size of the sample and the true population effect size. When looking at performance measures that take both bias and variability into account, such as the (root) mean squared error, none of the three indices is uniformly optimal and very little is known on in which situations one method outperforms another. The first goal of the current manuscript is to provide practical guidelines on how to deal with these different effect size estimates when used in a-priori power analysis based on the effect size estimate in a previous study.

3. Bias in power analyses

The sampling distributions of η^2, ω^2 and ε^2 are considerably skewed (shown in Fig. 1 for η^2). Furthermore, the smaller the sample size, the more variable the effect size estimate is. Statisticians have warned against using effect size estimates from small samples in power analyses (Leon, Davis, & Kraemer, 2011). The two main reasons researchers should be careful when using effect sizes from the published literature in power analyses is that effect size estimates from small studies are inaccurate, and that publication bias inflates effect sizes. At the same time, many applied statistics texts recommend using effect sizes from related studies reported in the literature to perform a power analysis (e.g., Fritz, Morris, & Richler, 2012; Polit & Beck, 2004; Sawyer & Ball, 1981). In many cases, this is the only information researchers have about the possible size of the effect they are interested in. For example, the Reproducibility Project (Open Science Collaboration, 2015) relied on the effect sizes observed in the original studies to perform power analyses for replication studies.

The statistical power of a test depends on the true effect size, the sample size, and the alpha level that is used. The goal of a power analysis is to control Type II error rates, or to limit the probability of observing a non-significant result, assuming there is an effect of a specific size. In the presence of bias, researchers might unknowingly increase the Type II error rate of their studies. Alternatives to a-priori power analysis exists, such as deciding upon a smallest effect size of interest and using this to determine the required sample size in a power analysis (e.g. Lakens & Evers, 2014, Lang & Sesic, 2006, denoted the 'minimal clinically important difference' in medical research, Jäschke, Singer, & Guyatt, 1989). Other researchers have suggested to perform conservative power analyses (Perugini, Gallucci, & Costantini, 2014), or to model and correct for bias (Taylor & Muller, 1996).

Nevertheless, researchers might believe that building on effect size

[4] Note that although these three indices are estimators, we adopt the usual convention to denote them without a 'hat'.

estimates from their own pilot studies is a valid approach, given that these effect size estimates are not influenced by publication bias. In this article, we show that even without problematic research practices such as publication bias or *p*-hacking, designing studies by relying on the effect size from a pilot study when performing an a-priori power analysis will, on average, lead to underpowered designs. As our simulations reveal, the deviation from the desired Type II error rate can be quite substantial, showing that the use of effect sizes from pilot studies will generally not be a good approach to designing future studies.

4. Follow-up bias

It is not uncommon that researchers perform a pilot study and use the information from the pilot study to decide whether or not to carry out a large scale follow-up study. The effect size in a pilot study is often used to determine what the sample size in a follow-up study should be based on an a-priori power analysis. This practice has been observed (and criticized) by statisticians (e.g., Kraemer, Mintz, Noda, Tinklenberg, & Yesavage, 2006), but as members of ethical review boards and local research funding committees, both authors often see that effect sizes used in power analyses are exclusively based on pilot data. According to Wald (1945), this double sampling inspection procedure dates back to Dodge and Romig (1929).

The likelihood that researchers will perform a follow-up study after a pilot study depends on how large the observed effect is, either expressed as a *p*-value, or a standardized effect size (for any given sample size, the effect size and the *p*-value for a statistical test are directly related). Researchers might decide to follow up on a study when the *p*-value is small (e.g., $p < 0.15$) or equivalently, when the effect size is large enough (e.g., $\eta^2 > 0.05$). When a pilot study yields a very small effect size estimate (or a very large *p*-value), power analysis will suggest a sample size is needed that is so large that the study is unfeasible, given limited resources. It could even be, when calculating ω^2 or ε^2, that the estimated effect size is negative, in which case a power analysis cannot be performed based on the observed effect size. Although researchers might in principle be interested in the presence vs. absence of an effect for purely theoretical reasons, in scientific practice they often have a maximum number of participants they are willing to collect for a study, either due to monetary or time constraints. Given any maximum sample size, there is a corresponding smallest effect size of interest (SESOI) that can be investigated with a decent level of power. Whenever effect size estimates in pilot studies are smaller than the SESOI, researchers might not to follow up on a line of research because they either suspect there is no true effect, or because examining this effect would require too much resources, if the effect size estimate in the pilot study is accurate. If researchers only follow-up on studies with a high η^2 estimate (e.g., $\eta^2 > 0.05$, which requires 76 participants in each condition) the effect sizes used in follow-up studies are on average upwardly biased, which leads to underpowered studies.

For any true effect size the variation in the effect size estimate from a pilot study can lead to *follow-up bias*: The tendency of researchers to not follow up on pilot studies where effect size estimates are small (e.g., those estimates close to 0 in Fig. 1), whereas the researcher would have followed-up on the pilot study if the true effect size had been known and accurately estimated. Suppose that a researcher has to decide between studying the effects of intervention A or B, but does not have the means to study both. Based on two pilot studies the researcher will conclude that the effect of one intervention, e.g. A, is larger than intervention B, and follow-up on Intervention A. This approach has recently been advocated by Sakaluk (2016). However, in small pilot studies, differences between effect sizes will themselves often not be statistically significant, and effect size estimates have high variability.

Both follow-up bias and publication bias lead to a preference for inflated effect sizes, but are not the same: Whereas follow-up bias is created by the boundaries a researcher sets for her/himself – most notably the maximum sample size they can collect – publication bias is introduced through the judgement of others (other researchers, reviewers, or editors). Researchers themselves are more likely to continue research lines where pilot studies estimate a larger compared to smaller effect size estimate, because the a-priori power analysis will indicate less resources are required to examine the effect with sufficient power. Therefore, whenever researchers choose to perform a study after performing a power analysis based on an observed effect size, an implicit selection process has already taken place, in that the follow-up study would not have been performed, had the pilot study revealed a tiny or even negative effect size estimate. By only performing follow-up studies when the observed effect size estimate falls on the right side of the distribution in Fig. 1, the effect size estimate that is used in power analyses is upwardly biased. Statistically, both publication bias and follow-up bias mean the power analysis is based on a truncated effect size distribution (Taylor & Muller, 1996).

In the first simulation study, where we aim to quantify the bias in power analysis due to the choice of the effect size index, we follow the procedure outlined in Appendix C. Power analysis for an ANOVA requires an iterative approach (outlined in Appendix B). In this paper, we apply this computational procedure as programmed in the *R* (*R* Core Team, 2017) package *pwr* (version 1.1–3, Champely, 2015) to examine how well the desired power is achieved as a function of the chosen effect size index and the maximum sample size a researcher is willing to collect, denoted by n^*.

In practical scenario's, it is impossible to distinguish the bias due to choice of effect size index from the follow-up bias; only the combined bias will be observed. Therefore, we start with a somewhat unpractical scenario, where the follow-up bias is practically zero. Thus, all observed bias will be due to the choice of effect size index. Once this first simulation study provides insight in how this type of bias operates, we will consider realistic scenario's and study the added bias due to follow-up bias in the second simulation study.

We first report the results of the simulation when using $n^* \geq 100{,}000$ as 'unpractically large to follow up on'. This large value means the follow-up bias in the simulation will be minimal, and allows us to focus on the consequences of different effect size calculations η^2, ω^2 and ε^2. Subsequently, we examine the consequences for follow-up bias given that most researchers have a maximum sample size per condition they are willing to collect that is substantially lower than 100,000. There, we employ values for n^* that are more in line with realistic situations in psychology.

5. Simulation design

For the simulation design, we varied the five parameters. First, in line with Okada (2013), the number of groups (*K*) in the ANOVA was 2, 3, or 4. This implies the simulations include the *t*-test and One-Way ANOVA with 3 and 4 groups. Second, the number of observations per group in the pilot study (n_{pilot}) was either 10, 25 or 50. A minimum of 50 participants in each condition for the pilot study has been recommended (Harris, 2001; Simmons, Nelson, & Simonsohn, 2013), but smaller sample sizes are still common (Fraley & Vazire, 2014). Third, the true population effect size (ES_p) was small (.0099), medium (.0588), or large (.1379), in line with Cohen's (1988) rules of thumb and corresponding to using $d = .2, .5,$ or .8, respectively. These values are also consistent with Okada's (2013) simulation design. Fourth, the desired power in the follow-up study, (0.9 or 0.8 power, i.e. $\beta = 0.1$ or 0.2). Finally, for the pilot and follow-up studies we calculated three effect size indices from the sample (ES_s) used to estimate ES_p, namely η^2, ω^2 or ε^2.

For all simulations, the significance level was set at $\alpha = 0.05$. We decided not to vary this parameter of the design as people almost exclusively work with $\alpha = 0.05$, and this would thus unnecessarily increase the complexity of the simulation design. All five parameters were fully crossed, yielding $3 \times 3 \times 3 \times 3 \times 2 = 162$ combinations. For each combination, $R = 1{,}000{,}000$ replications were drawn (the

supplementary material shows this number to be sufficient); yielding a total of 162 million simulated samples. A detailed breakdown of the steps in the simulation design is provided in Appendix C and the corresponding *R* script is available at https://osf.io/zq9mg/.

6. Results

Consider a study designed to test a difference between two groups (so, $K = 2$, a *t*-test). We will first focus on a true population effect size that is large ($ES_p = .1379$), before investigating small and medium effects. A pilot study, consisting of n_{pilot} measurements for each group, is performed and the data from this sample is used to estimate the population effect size (as long as the effect size estimate in the pilot study is larger than 0 for ω^2 and ε^2). Power analyses are performed to compute the sample size of the main study to reach a power of 0.8, assuming the effect size estimate from the pilot study is the true population effect size. Subsequently, a data set of this size suggested by the power analysis is generated, and the true power of the main study is calculated. Note that in practice, the true population effect size and the true power of studies is unknown, and these values are only known in simulation studies.

Fig. 2 shows box plots and violin plots for effect sizes calculated from the results of the main study, for all three effect size indices, based on pilot studies with n_{pilot} 10, 25, or 50 participants (left facet, middle facet, and right facet) in each condition. The simulation confirms that the mean estimates (indicated by the black dots in Fig. 2) are very close to the true population effect size of .1379 (indicated by the dashed horizontal lines), with η^2 having a (relatively small) positive bias. In line with previous work (Okada, 2013), ω^2 and ε^2 give mean effect size estimates closer to the true effect size. The average bias (or deviation from the true effect size) is provided in detail in Table 1.

Furthermore, it can be seen in Fig. 2 that that when n_{pilot} increases (e.g., right facet), the variation in the estimates decreases. This is not surprising: With a larger sample, estimates will be, on average, more accurate. Especially with small pilot samples, it is likely that the effect size is either severely under- or overestimated. This, in turn, leads to sample sizes for the main study that are either too low or too high. Thus, small pilot studies lead to large variation in effect size estimates both in the pilot study itself, as in the follow-up study when the sample size is based on the effect size observed in the pilot study.

There is another complication in determining the sample size of the main study, n_{main}. As Fig. 1 shows, the relation between estimated effect size in the pilot and n_{main} following a power analysis is clearly non-linear. Because of the uncertainty in estimating the effect sizes in small samples, effect sizes estimates from the samples will vary around the true effect size. A higher estimate for the effect size will yield a lower value for n_{main} in a power analysis, and a lower estimate for the effect size will yield a higher value for n_{main}. The median of the effect size distribution lies roughly at $\eta^2 = 0.05$, which corresponds to $n_{main} = 76$ measurements per group in the main study. An estimated effect size 0.025 above this median, yields $n_{main} = 50$, whereas an estimated effect size of 0.025 below this median yields $n_{main} = 155$ Clearly, 50 is much closer to the accurate sample size required to achieve 0.8 power of 76 observations per group than 155 is. Due to the non-linear relation between sample size and power, the probability that n_{main} is *severely* overestimated is much larger than the probability that n_{main} is *severely* underestimated. In mathematical statistics, this phenomenon is known as Jensen's inequality (Jensen, 1906).

This is demonstrated in Fig. 3. The distribution of n_{main} based on an a-priori power analysis (assuming the effect size in the pilot study is the true effect size) is very skewed, showing that, on average, the sample size for the main study is considerably overestimated. Especially with very small sample sizes, this leads to follow-up studies where the median sample size falls below the required sample size to reach a power of 0.8.

As a final and most important step, we can see the consequences of the too small sample sizes in the main study when we analyze the power of the main study. Fig. 4 shows that with only $n_{pilot} = 10$ measurements per

Table 1
Mean bias in effect size estimate for the 3 × 3 combinations of population effect size and pilot sample size, for the three effect size measures.

	Population effect size				
	n_{pilot}	Small	Medium	Large	Average
η^2	10	+ 0.0153	+ 0.0232	+ 0.0349	+ 0.0245
	25	+ 0.0068	+ 0.0150	+ 0.0270	+ 0.0163
	50	+ 0.0042	+ 0.0124	+ 0.0245	+ 0.0137
	Average	+ 0.0088	+ 0.0169	+ 0.0288	+ 0.0182
ε^2	10	− 0.0001	− 0.0005	− 0.0016	− 0.0007
	25	− 0.0000	− 0.0004	− 0.0013	− 0.0006
	50	− 0.0000	− 0.0003	− 0.0012	− 0.0005
	Average	− 0.0000	− 0.0004	− 0.0014	− 0.0006
ω^2	10	− 0.0001	− 0.0012	− 0.0034	− 0.0016
	25	− 0.0001	− 0.0007	− 0.0029	− 0.0012
	50	− 0.0000	− 0.0006	− 0.0028	− 0.0012
	Average	− 0.0001	− 0.0008	− 0.0030	− 0.0013

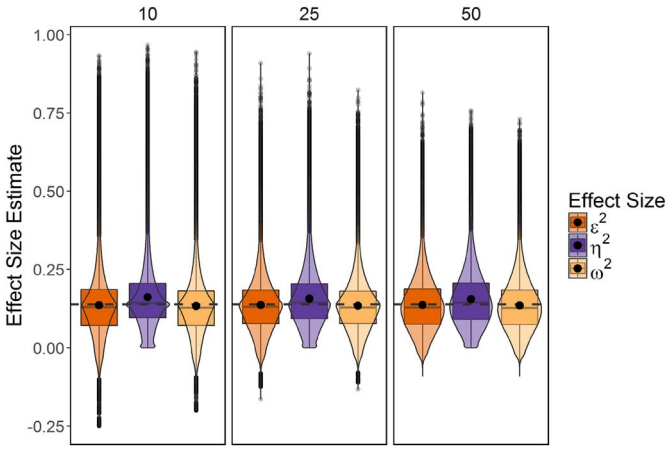

Fig. 2. Box plots (marking the median and the 1st and 3rd quartiles of the distribution) and violin plots for effect size estimates for *t*-tests with a large true population effect size. The dashed line marks the true effect size, and the dots mark the mean effect size estimate in the simulation. The panels denote, from left to right, n_{pilot} values of 10, 25, and 50. Reprinted from http://dx.doi.org/10.6084/m9.figshare.5198050, CC-BY4.0.

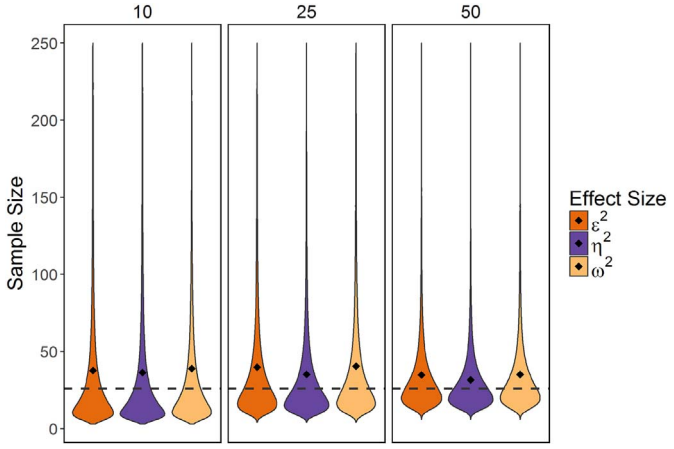

Fig. 3. Box plots and violin plots for the estimated sample size per condition required to reach 0.8 power based on the effect size estimate from a pilot study with n_{pilot} (left: 10, middle: 25, right: 50) participants per condition for the *t*-test with a large population effect size (0.1379). The dashed horizontal line indicates the required sample size ($n = 26$) to achieve 0.8 power for the true effect size. The vertical axis is capped at $n = 250$. Reprinted from http://dx.doi.org/10.6084/m9.figshare.5198053, CC-BY4.0.

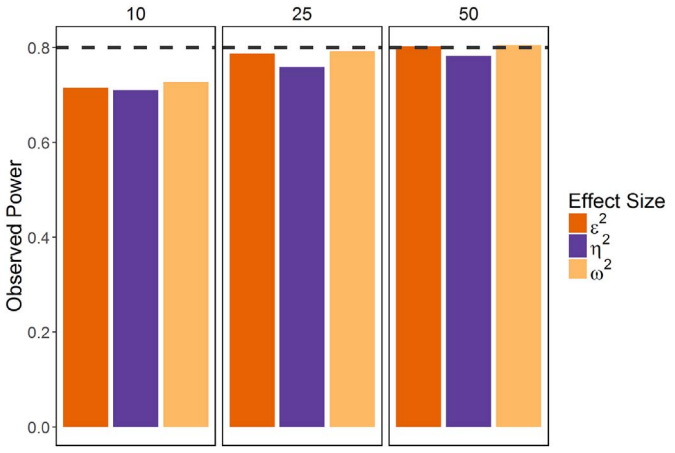

Fig. 4. Mean power for a *t*-test, when the sample size for the main study is based on an a-priori power analysis to achieve a power of 0.8 (dotted line) based on the effect size estimate observed in a pilot study with n_{pilot} (left: 10, middle: 25, right: 50), when the true effect size is large (.1379). Reprinted from http://dx.doi.org/10.6084/m9.figshare.5198056, CC-BY4.0.

group, the main study is seriously underpowered, on average. The reason for this is mainly the *accuracy bias* due to the skewness introduced by Jensen's inequality.

6.1. Results for lower population effect sizes

So far, we have looked at simulations where the population effect size is large. Power analyses based on relatively small pilot studies become more inaccurate when the effect size is medium or small. Fig. 5a displays the true power for follow-up studies when examining small effect sizes, and Fig. 5b displays the true power for medium effect sizes, complementing Fig. 4.

The smaller the sample size of the pilot study, the larger the variance of the effect size estimate, and thus the wider its distribution. For ε^2 and ω^2, this means that a relatively larger percentage of effect size estimates falls below 0, and cannot be used for a power analysis. When power analyses are performed using ε^2 and ω^2, the effect sizes calculated from the pilot studies overestimate the true effect size as explained above, leading to relatively underpowered studies. Fig. 5a shows that for small effect sizes and $n_{pilot} = 10$, the main study is extremely underpowered with power less than half of what it should be. Perhaps situations like these won't occur often in practice: even with an alpha-level of 0.10 or 0.15, pilot studies are likely to be non-significant and interest in the line of research would diminish. When true effect sizes are small, using effect sizes estimates from small pilot studies to perform power analyses for a main study is not a useful approach to design well-powered main studies. Main studies remain underpowered, but to a lesser extent, when n_{pilot} is set to 25, or with $n_{pilot} = 50$. The observed power is closer to the true power in these studies. Furthermore, it can be seen that the differences between the three effect sizes, η^2, ω^2 and ε^2, are substantial. Even though η^2 is positively biased, the fact that it can't be negative yields follow-up studies with better power (because more power analyses yield sample size estimates below 100,000), but only when effect sizes are small, and/or when effect sizes are medium, and pilot studies are small (situations that might not often occur in practice).

7. Follow-up bias

In the simulations reported above, we have set the maximum sample size for a follow-up study to 100,000 participants to be able to illustrate the effects of accuracy bias. With extremely rare exceptions, collecting 100,000 participants will not be feasible in practice. Fraley and Vazire (2014) examined the total sample sizes in six psychology journals between 2006 and 2010, and found that median total sample sizes range from 211 to as low as 51.5. It is therefore important to examine the consequences of follow-up bias, or the tendency to only perform follow-up studies when sample sizes in pilot studies are sufficiently large.

We can simulate the consequences of follow-up bias as a function of the maximum sample size a researcher is willing, or has the resources, to collect. Remember that for a given study, the choice to perform a follow-up study based on a maximum sample size is directly related to the smallest effect size or a largest *p*-value a researcher will decide to follow up on (see Fig. 1). By re-analyzing our simulation results, we can examine what happens when the maximum sample size in a follow-up study is lowered from 100,000 to more realistic values. In Fig. 6 we present these re-analyses for small, medium, and large true effect sizes. We only look at η^2 for sake of simplicity (the effects of follow up bias on ω^2 and ε^2 being similar to η^2), and examine follow-up studies designed to have a power of 0.8 and a pilot study with $n_{pilot} = 25$.

The power when follow-up studies have a maximum sample size of 100,000 is 0.58, 0.72, and 0.76 (the middle bar in Figs. 4 and 5a,b, and the dashed lines in Fig. 6). When the maximum sample size a researcher is willing to collect lies below 250 participants per group, the true power in follow-up studies is even lower. For example, when examining a medium true effect size (the middle red line in Fig. 6), and relying on a pilot study with 25 participants in each sample, researchers who are willing to collect a maximum of 100 participants in each of two groups (so 200 in total) will achieve at most a power of 0.6, in the long run. For

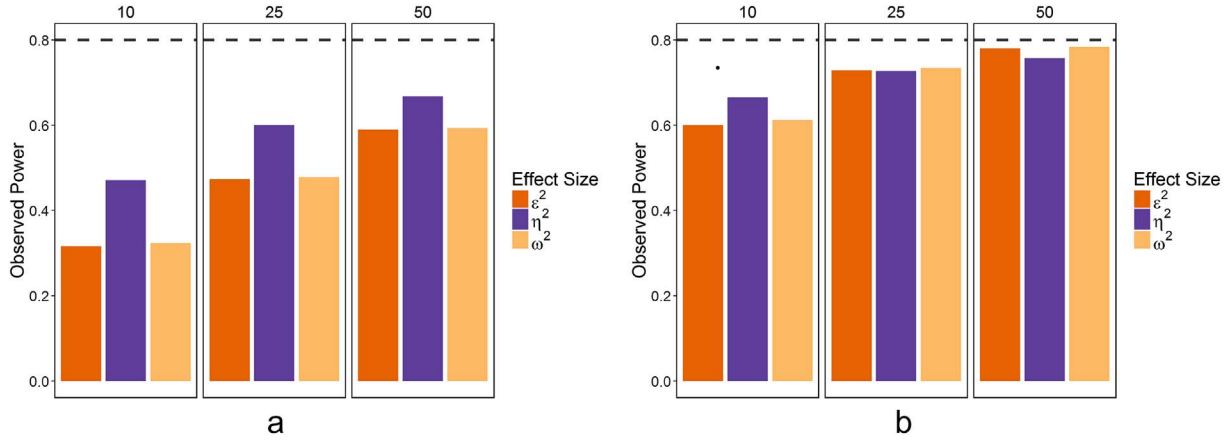

Fig. 5. a (left) and b (right). Mean power for a *t*-test, when the sample size for the main study is based on an a-priori power analysis to achieve a power of 0.8 (dotted line) based on the effect size estimate observed in a pilot study n_{pilot} (left: 10, middle: 25, right: 50), when the true effect size is small (.0099; panel a) and medium (0.0588; panel b). Reprinted from http://dx.doi.org/10.6084/m9.figshare.5198059 and http://dx.doi.org/10.6084/m9.figshare.5198062, CC-BY4.0.

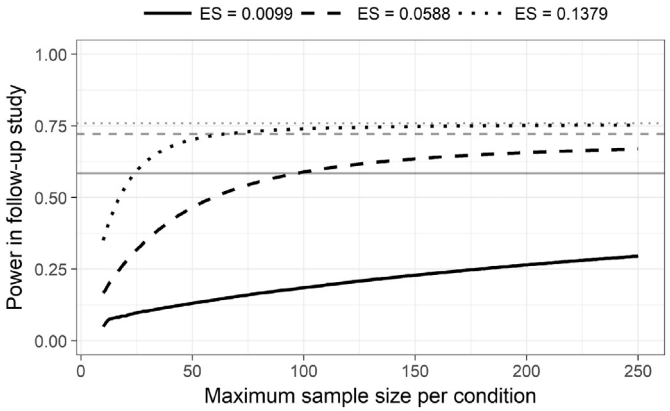

Fig. 6. Power for main studies (where the sample size is based on a power analysis with an effect size estimated from a pilot study with $n_{pilot} = 25$), as a function of the true effect size and the maximum sample size per group a researcher is willing to collect in the main study. Reprinted from http://dx.doi.org/10.6084/m9.figshare.5198065, CC-BY4.0.

large true effects, the additional bias in the true power due to follow-up bias is only pronounced when the maximum sample size a researcher is willing to collect is small (see the upper curve in Fig. 6). For small true effects, follow-up bias leads to main studies that are not even close the desired 0.8 power. Fig. 6 clearly shows that designing main studies based on effect sizes estimated from small pilot studies does not yield desired levels of power due to accuracy bias and follow-up bias. Tables S4 to S6 in the Supplementary Material provide the observed power for each of the 162 conditions for various levels of n^*.

7.1. More than two groups

When the number of groups increases, the differences in power and the proposed sample size for the main study due to differences in n_{pilot} become somewhat smaller: The biases we discuss in this paper are more severe for the *t*-test than for a four group ANOVA. This, however, is largely due to the set-up of the simulation study, where an increase in the number of groups also mean an increase in the total sample size. For example, $n_{pilot} = 25$ means that the total sample size equals $K \times 25$. For a *t*-test this is $2 \times 25 = 50$, whereas for a one-way ANOVA with four groups it is $4 \times 25 = 100$. For instance, with two groups and a large effect size, a pilot study with a total sample size of 50 has a power of 0.79, but with three groups and a large effect, the pilot study has a total sample size 75, and 0.87 power. The Supplementary Material provides similar information for each of the 1628 conditions. Tables S1 to S3 provide, for the three types of effect size studied in this paper, the median n_{main}, the average and median bias, the root mean-squared error, the percentage of simulations where the power analysis yielded a sample size estimate above n^* and the average achieved power.

7.2. Change in power

When the power is set to 0.9 rather than 0.8, the impact of negative or small effect size estimates is roughly similar. Averaged over all conditions with 0.8 desired power, the achieved power is 0.647, which is 80.9% of 0.8 power. Averaged over all conditions with 0.9 desired power, the achieved power is 0.717, which is 79.7% of 0.9 power. When the true population effect is small, and/or the pilot study is small, the observed power is slightly better when studies were designed to have 0.9, compared to 0.8 power. In the other conditions, the opposite holds. (see Tables S1, S2, and S3).

7.3. Summary

Tables 2 summarizes the findings of Tables S1, S2 and S3 for the different values of one parameter, averaging across all other parameters in the simulation, with respect to K, n_{pilot}, ES_s and ES_p. Overall, η^2 yields considerably less powerful main studies than ω^2 and ε^2. Furthermore, in *all* situations, the average power lies *well* below the desired level. On average, the achieved power is about 80% of the desired power. Out of all 162 simulated conditions, only nine conditions have an average power level that reaches (or exceeds by 1%) the desired level of power. These conditions aimed for a desired power of 0.8, examined a large true effect, with two conditions having $n_{pilot} = 25$, and the other six have $n_{pilot} = 50$. None of the seven conditions relied on η^2. Ironically, these are exactly the situations least representative of current practices in psychology, where effect sizes are often not large (Richard, Bond, & Stokes-Zoota, 2003), studies have low sample sizes (Fraley & Vazire, 2014), and η^2 is used more often than ε^2 or ω^2 (Open Science Collaboration, 2015).

8. Discussion

We have shown that the practice of conducting a pilot study, estimating the effect size from the data, and using the effect size estimate in an a-priori power analysis to decide upon the sample size of the follow-up study leads to substantially underpowered main studies in most realistic situations. Although researchers are often reminded that effect size estimates from small studies can be unreliable (e.g., Lakens & Evers, 2014), researchers are rarely informed about the consequences of using biased effect size estimates in power analyses. Researchers who design studies based on effect size estimates observed in pilot studies will unknowingly design on average underpowered studies, as long as they don't take bias in the estimated effect sizes and follow-up bias into account. The difference between the desired and achieved power can be especially worrying when the sample size of the pilot study and/or the population effect size is small, or when researchers are not willing to collect large sample sizes in the main study. It is important that researchers are aware of these pitfalls when designing studies.

Based on the results of the simulations in this manuscript, we offer the following recommendations for researchers who want to decide upon a sample size when designing their study. First and foremost, we recommend against performing a pilot study to estimate an effect size, and subsequently using this effect size estimate in a power analysis to design a follow-up study. This can lead to seriously underpowered study designs, especially when the sample size of the pilot and/or the true effect size is small to medium. Not only is this approach inaccurate, it is inefficient. The data from the pilot study is either completely ignored, or at least not included in the main study. We don't see how this waste of pilot data can be justified, when the end result is a procedure that yields underpowered main studies. Pilot studies can be performed because they have other uses, such as determining the feasibility of performing the designed study in practice, which is especially useful when trying out new methods or procedures (Leon et al., 2011).

Alternative approaches exist. First, one can determine the smallest effect size of interest (SESOI), based on either utility or theoretical arguments, and use the SESOI in an a-priori power analysis. This leads to main studies that have a pre-determined statistical power to detect or reject the smallest effect size that is deemed worthwhile to study. For example, if researchers decide their SESOI is a medium effect size of $\eta^2 = .0588$ a study with 87 participants in each of two groups will in the long run have a power of 0.9 to detect the SESOI, or reject it in an equivalence test (Lakens, 2017). Choosing a SESOI allows researchers to control their Type II error rate exactly for effect sizes they care about.

Alternatively, researchers might simply decide upon the maximum sample size they are willing to collect. Such a maximum number can and should be motivated, e.g. based on theoretical constraints or the amount of means available. Fig. 7 shows the statistical power as a function of the sample size for a two-group ANOVA for small, medium, and large effects. Simply collecting a maximum number of observations will lead to considerably higher power than aiming for a power of 0.8, given a maximum sample size you are willing to collect (see Fig. 6), but

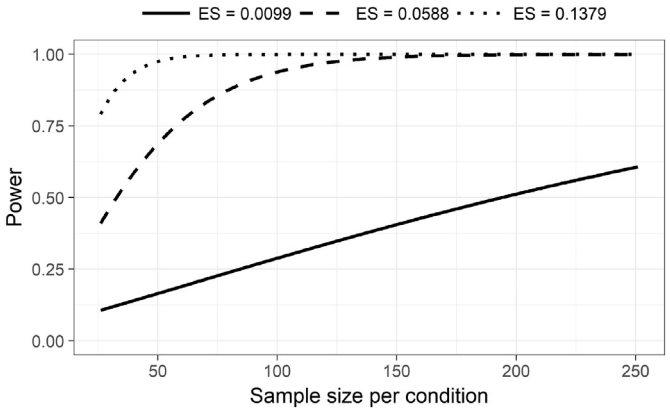

Fig. 7. Power for a two-group ANOVA as a function of the true effect size and the sample size per group. Reprinted from http://dx.doi.org/10.6084/m9.figshare.5198068, CC-BY4.0.

it is less efficient (you will end up collecting more participants than when you had performed an a-priori power analysis).

A more efficient alternative approach to designing studies is to use sequential analyses, which allows researchers to analyze the data multiple times (e.g., after 50, 100, 150, and 200 participants have been collected) whilst controlling Type I error rates. For a frequentist introduction of sequential analysis, see Lakens (2014), for a Bayesian introduction, see Schönbrodt, Wagenmakers, Zehetleitner, and Perugini (2017), for the mathematical background, see Wald (1945), and Siegmund (2013). It can be shown mathematically that sequential analyses are more efficient than the double sampling scheme with a pilot study, and sequential analyses are especially appropriate whenever the true effect size is relatively uncertain. However, sequential designs are not always feasible. For example, in a 5-year longitudinal study, one can only look at the data after five years, and each subsequent look at the data adds another 5 years to the research project. In such designs, power analyses will still be an important aspect of the study design. In such situations, and alternative approach is to perform a *conditional power analysis*, where an initial *internal* pilot study is collected, based on which a power analysis is performed (if the effect is not yet significant based on the available data), after which the remainder of the required sample is collected, and all data is combined in the final analysis.

We believe recent recommendations such as "Explore small, confirm big" (Sakaluk, 2016) are not useful. Small pilot studies only provide useful information to design follow-up studies when effect sizes are large. When the true effect size is large ($\eta^2 = 0.1379$) even relatively small follow-up studies (e.g., $n = 34$ in each group for a *t*-test) have sufficient (i.e., 0.9) power. When the true effect size is small or medium, small pilot studies will have low power and effect size estimates have high variability, which make it difficult to decide whether, or how, a follow-up study should be designed. Do not use extremely small sample sizes (e.g., $n_{pilot} = 10$) to estimate the true effect size. These are *too* small to get even remotely accurate estimates for n_{main} in a power analysis.

Do not use η^2 in power analyses as this leads to the lowest power, on average (Table 2) – use ε^2 or ω^2 instead. Note that 56% of the studies in the Reproducibility Project used η^2 or ${\eta_p}^2$ as the effect size index for the a-priori power analysis.

Power analyses based on pilot studies almost always yield estimates of the sample size of the main study n_{main} that are too low. Whenever power analyses are based on effect size estimates from previous research (either pilot studies, or published studies) we recommend researchers take measures to compensate for this bias. A possible solution is to perform power analyses with $\tilde{\eta}^2$ rather than η^2, where $\tilde{\eta}^2$ is the lower bound of a 80% confidence interval for η^2. This recommendation, known as safeguard power analysis, was proposed by Perugini et al. (2014). Alternatively, one can model the bias, and calculate n_{main} based on a truncated *F*-distribution. This approach, building on work by Taylor and Muller (1996), was recently recommended by Anderson and Maxwell (2017). Researchers can choose the level of truncation (e.g., making the assumption that only studies with $p < 0.05$ appear in the literature), and perform a power analysis based on this truncated *F*-distribution. When sequential analyses are not possible, the use of safeguard power or truncated *F*-distributions are good approaches to compensate for the bias in traditional power analyses where the effect size is derived from a pilot study or the published literature.

Table 2
Average estimates of the observed power in main studies for a desired power level of 0.8 and 0.9 for (i) number of groups, (ii) the size of pilot study, (iii) effect size index, (iv), and true population effect size.

	0.8	0.9
$K = 2$	.659	.726
$K = 3$	.644	.715
$K = 4$	.638	.711
$n_{pilot} = 10$	.553	.617
$n_{pilot} = 25$	.665	.736
$n_{pilot} = 50$	.723	.799
ε^2	.655	.724
η^2	.626	.699
ω^2	.660	.728
ES_p small	.472	.534
ES_p medium	.705	.777
ES_p large	.764	.840

8.1. Suggestions for future research

This manuscript focused on the use of effect sizes to determine the sample size in follow-up studies using balanced one-way ANOVA's. For unbalanced designs (designs with substantially different sample sizes per cell) the bias in power analysis might be different (Kline, 2013). Furthermore, in other experimental designs, such as regressions, more-way factorial ANOVAs and within-subject designs, follow-up bias will also distort power analyses. The extent of the bias in these designs could be quantified in future simulation studies. Researchers interested in preventing bias in those situations are recommended to adapt our simulations to their situation of interest, or use sequential designs instead. Furthermore, because the strength of the bias in power analysis is most severe if the maximum sample size researchers are willing to collect is relatively small, and the SESOI is thus relatively large, it is interesting to empirically examine what the distribution of the SESOI and maximum sample size researchers are willing to collect is in different research domains, and how this affects follow-up bias.

Researchers become increasingly aware of the importance of designing well-powered studies. Several journals now require authors to justify their sample sizes, and although power analyses are only one possible justification, it seems likely power analyses will become more widely used. Power analysis can be one of the factors informing a study design, but it should not be mechanistically used to determine the sample size that will be collected. We feel it is important that researchers realize possible sources of bias in the estimated sample sizes they need for a desired level of power, and are advised to attempt to correct for these biases when designing a study, or use other approaches to determine the sample size of a study.

Open practices

The study in this article earned the Open Materials badge for transparent practices. Materials for this study are available at https://osf.io/zq9mg/.

Acknowledgements

We wish to thank the reviewers, Dr. Marco Perugini, Michèle Nuijten, MSc., and an anonymous reviewer, for their comments that helped to improve the manuscript considerably.

Appendix A. Computation of effect sizes based on reported *F*-statistics

The formulas on page 5 seem to suggest that full details of the ANOVA table are required to compute the effect sizes. This is not the case: with some algebra (Carroll & Nordholm, 1975; Cohen, 1988) it can be shown that, in the between-subject designs studied in this paper, all effect sizes can be computed on basis of reported F values and degrees of freedom only. (Note that for the t-test, $F = t^2$ and $df_B = 1$).

In a one-way ANOVA η^2 equals ${\eta_p}^2$ (and $\omega^2 = {\omega_p}^2$, and $\varepsilon^2 = {\varepsilon_p}^2$). In more complex designs than One-Way ANOVA's partial effect sizes are used in a-priori power analysis. Researchers often don't report ${\omega_p}^2$ or ${\varepsilon_p}^2$. Fortunately, for designs where all factors are manipulated (but not for studies with measured factors or ANCOVA's) these (partial) effect size indices, as well as Cohen's f, can be calculated from the F-value and both degrees of freedom:

$$\eta_p^2 = \frac{F \times df_B}{F \times df_B + df_W},$$

$$\omega_p^2 = \frac{F - 1}{F + \frac{df_W + 1}{df_B}},$$

$$\varepsilon_p^2 = \frac{F - 1}{F + \frac{df_W}{df_B}},$$

$$f = \sqrt{F \times \frac{df_B}{df_W}}.$$

A spreadsheet document to calculate ${\eta_p}^2$, ${\omega_p}^2$ and ${\varepsilon_p}^2$ from the F-value and degrees of freedom is available from https://osf.io/zq9mg/.

Appendix B. Computational power analysis

Power analyses are computationally tricky and therefore usually performed using software as G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) or the *pwr R* package (Champely, 2015). For a One-Way ANOVA, it requires an iterative approach, which is as follows (cf. Champely, 2015; Cohen, 1988; Faul et al., 2007). First, the so-called non-centrality parameter λ is specified via

$$\lambda = \frac{n\delta^2}{2MS_W},$$

where n denotes the required sample size per group and δ^2 the effect size in the population. Next, n is obtained by equating.

$$F_{\alpha,k-1,n(k-1)} = F'_{1-\beta,k-1,n(k-1),\lambda} \quad (1)$$

Here, α denotes the level of significance and $1 - \beta$ denotes the desired power (cf. Cohen, 1988). A solution for n is found through the bisection method:

(i) Specify a lower bound n_* and upper bound n^* for n,
(ii) Compute Eq. (1) for $n' = \frac{n_* + n^*}{2}$,
(iii) When the left-hand-side of (*) exceeds the right-hand-side set $n_* = n'$, else set $n^* = n'$,

Repeat steps (i)–(iii) until the absolute difference between the left-hand side and right-hand side of Eq. (1) is smaller than some pre-specified tolerance level.

Appendix C. Overview of the six steps in the simulation study

1. Input: K, n_{pilot}, ES_p, ES_s, β, and α.
2. Compute $f = \sqrt{(ES_p/1 - ES_p)}$
3. Generate pilot data: n_{pilot} observations per group, from normal distributions with means: $-f$ and $+f$ ($K = 2$), $-f$, 0 and $+f$ ($K = 3$), $-f$, $-f$, $+f$ and $+f$ ($K = 4$) and a standard deviation of 1.
4. Based on the pilot data, estimate either η^2, ω^2 or ε^2.
5. Perform an a-priori power analysis using the *pwr* package to determine the sample size per group of the main experiment. When this exceeds $n^* = 100{,}000$ (e.g. when $\omega^2 = 0$), the sample size is coded as '*not available*'.
6. For all performed power analyses, a study identical to the pilot study but with a sample size based on the power analysis was performed. For these studies, we calculated three effect size estimates (η^2, ε^2, and ω^2), performed the statistical test and stored the p-value and the sample size of the study. When the power analysis in step 5 yielded NA, all values for the effect size, p-value, and sample size were set to NA as well.

Appendix D. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jesp.2017.09.004.

References

Albers, C. J. (2015). *Comment on "why you should use omega^2 instead of eta^2".* (June 10). [Web log post]. Retrieved from http://blog.casperalbers.nl/science/statistics/comment-on-why-you-should-use-omega%C2%B2-instead-of-eta%C2%B2/.

Anderson, S. F., & Maxwell, S. E. (2017). Addressing the "replication crisis": Using original studies to design replication studies with appropriate statistical power. *Multivariate Behavioral Research,* 1–20. http://dx.doi.org/10.1080/00273171.2017.1289361.

Carroll, R. M., & Nordholm, L. A. (1975). Sampling characteristics of Kelley's ε and Hays' ω. *Educational and Psychological Measurement, 35*(3), 541–554.

Champely, S. (2015). Pwr: basic functions for power analysis. *R package version, 1*, 1–3 Retrieved from http://CRAN.R-project.org/package=pwr.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). Mahwah, New Jersey: Lawrence Erlbaum.

Dodge, H. F., & Romig, H. G. (1929). A method of sampling inspection. *Bell Labs Technical Journal, 8*(4), 613–631. http://dx.doi.org/10.1002/j.1538-7305.1929.tb01240.x.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191. http://dx.doi.org/10.3758/BF03193146.

Fisher, R. A. (1928). *Statistical methods for research workers* (2nd ed). Edinburgh, UK: Oliver and Boyd.

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), e109019. http://dx.doi.org/10.1371/journal.pone.0109019.

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.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82*(1), 1–20.

Harris, R. J. (2001). *A primer of multivariate statistics* (3rd edition). Oxford: Psychology Press.

Hays, W. L. (1963). *Statistics for psychologists.* New York: Holt: Rinehart and Winston.

Jäschke, R., Singer, J., & Guyatt, G. H. (1989). Measurement of health status: Ascertaining the minimal clinically important difference. *Controlled Clinical Trials, 10*(4), 407–415. http://dx.doi.org/10.1016/0197-2456(89)90005-6.

Jensen, J. L. W. V. (1906). Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Math. 30*(1), 175–193. http://dx.doi.org/10.1007/BF02418571.

Kelley, T. L. (1935). An unbiased correlation ratio measure. *Proceedings of the National Academy of Sciences, 21*, 554–559. http://dx.doi.org/10.1073/pnas.21.9.554.

Kline, R. B. (2013). *Beyond significance testing: Reforming data analysis methods in behavioral research* (Second Edition). Washington, DC: American Psychological Association.

Kraemer, H. C., Mintz, J., Noda, A., Tinklenberg, J., & Yesavage, J. A. (2006). Caution regarding the use of pilot studies to guide power calculations for study proposals. *Archives of General Psychiatry, 63*(5), 484–489. http://dx.doi.org/10.1001/archpsyc.63.5.484.

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. http://dx.doi.org/10.3389/fpsyg.2013.00863.

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

Lakens, D. (2015, June 8). *Why you should use omega-squared instead of eta-squared [Web log post].* Retrieved from http://daniellakens.blogspot.nl/2015/06/why-you-should-use-omega-squared.html.

Lakens, D. (2017). Equivalence tests: A practical primer for *t* tests, correlations, and meta-analyses. *Social Psychological and Personality Science, 8*(4), 355–362. http://dx.doi.org/10.1177/1948550617697177.

Lakens, D., & Evers, E. (2014). Sailing from the seas of chaos into the corridor of stability: Practical recommendations to increase the informational value of studies. *Perspectives on Psychological Science, 9*, 278–292. http://dx.doi.org/10.1177/1745691614528520.

Lang, T. A., & Sesic, M. (2006). *How to report statistics in medicine: Annotated guidelines for authors, editors and reviewers* (2nd edition). Philadelphia: American College of Physicians.

Leon, A. C., Davis, L. L., & Kraemer, H. C. (2011). The role and interpretation of pilot studies in clinical research. *Journal of Psychiatric Research, 45*(5), 626–629. http://dx.doi.org/10.1016/j.jpsychires.2010.10.008.

Levine, T. R., & Hullett, C. R. (2002). Eta squared, partial eta squared, and misreporting of effect size in communication research. *Human Communication Research, 28*(4), 612–625. http://dx.doi.org/10.1111/j.1468-2958.2002.tb00828.x.

McGrath, R. E., & Meyer, G. J. (2006). When effect sizes disagree: The case of r and d. *Psychological Methods, 11*(4), 386–401. http://dx.doi.org/10.1037/1082-989X.11.4.386.

Okada, K. (2013). Is omega squared less biased? A comparison of three major effect size indices in one-way ANOVA. *Behaviormetrika, 40*(2), 129–147. http://dx.doi.org/10.2333/bhmk.40.129.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6521), aac4716. http://dx.doi.org/10.1126/science.aac4716.

Pearson, K. (1911). On a correction needful in the case of the correlation ratio. *Biometrika, 8*, 254–256.

Perugini, M., Gallucci, M., & Costantini, G. (2014). Safeguard power as a protection against imprecise power estimates. *Perspectives on Psychological Science, 9*(3), 319–332. http://dx.doi.org/10.1177/1745691614528519.

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

R Core Team (2017). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing ISBN 3-900051-07-0, URL http://www.R-project.org/.

Richard, F. D., Bond, C. F., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology, 7*(4), 331–363. http://dx.doi.org/10.1037/1089-2680.7.4.331.

Sakaluk, J. K. (2016). Exploring small, confirming big: An alternative system to The New Statistics for advancing cumulative and replicable psychological research. *Journal of Experimental Social Psychology, 6*, 47–54 doi: j.jesp.2015.09.013.

Sawyer, A. G., & Ball, A. D. (1981). Statistical power and effect size in marketing research. *Journal of Marketing Research, 18*, 275–290.

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods, 22*(2), 322–339. http://dx.doi.org/10.1037/met0000061.

Siegmund, D. (2013). *Sequential analysis: Tests and confidence intervals.* New York: Springer.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2013). *Life after P-hacking (SSRN scholarly paper no. ID 2205186).* Retrieved from http://papers.ssrn.com/abstract=2205186.

Taylor, D. J., & Muller, K. E. (1996). Bias in linear model power and sample size calculation due to estimating noncentrality. *Communications in Statistics-Theory and Methods, 25*(7), 1595–1610.

Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statstics, 16*(2), 117–186. http://dx.doi.org/10.1214/aoms/1177731118.