# Chapter 10. Approaches to effect size selection for sample size planning

## Abstract

This paper presents the results of an analysis of the sample size planning approaches used in 121 empirical research articles published in the November 2017 to August 2018 issues of Psychological Science and uses the results of this analysis to illustrate a guide to effect size selection for power analysis. When formal sample size planning is used it is important to understand that the method used to develop the alternative hypothesis (i.e., the effect sizes and parameter estimates used in power analysis) has important implications for the appropriate interpretation of the results of sample size planning. This paper outlines how different effect size selection methods impact the meaning of power analyses, discusses the appropriateness of each approach under different scenarios, the key difficulties in each approach, and how to address each of these issues. Finally, this paper introduces several less common methods of sample size planning (Accuracy in Parameter Estimation, Statistical Assurance, and Bayesian sample size determination). This paper provides guidance on how to select effect sizes under these different approaches to sample size planning, and explains the implications that follow from each effect selection method while reinforcing some important warnings against practices that are likely to lead to inaccurate inferences about the power or precision of planned research.

Keywords: Statistical power, effect size, sample size, power analysis

## Introduction

Researchers routinely have to decide upon the sample size they include in their research. When statistical tools are used to help guide sample size decisions in psychology, researchers typically use statistical power analysis (Cohen, 1988). Statistical power analysis allows researchers to show the probability with which a statistical test will reject the null hypothesis under a specified alternative hypothesis. An under-addressed issue in the literature discussing statistical power analysis is the question of how researchers should develop the alternative hypotheses they use in their power analyses. There are two main approaches to effect size selection that are typically discussed in the psychology literature, estimating power at the effect size (Anderson, Kelley, & Maxwell, 2017), and estimating power at a minimum interesting or clinically significant effect size (Biau, Kernéis, & Porcher, 2008). A third approach that is rarely discussed in the psychology literature is to use a Bayesian prior distribution over effect sizes and parameters. This can then be used either to estimate ‘assurance’, the probability that a study will find statistically significant results (Ren & Oakley, 2014), or to estimate the probability of developing convincing evidence or precise credible interval estimates if a Bayesian approach to statistical analysis is to be used.

Any formal sample size determination method posits a hypothetical scenario (or probability distribution over effect sizes) and is only meaningful with regards to the proposed parameter values or prior distribution. Estimating the expected effect size, the minimum effect of interest, or using a Bayesian prior distribution may be appropriate in different circumstances, although each provides very different information. This paper outlines these different approaches to developing alternative hypotheses in formal power analysis and explains the implications of each. It also briefly introduces alternatives to power analysis that may be more appropriate when researchers’ primary research goals are not to reject a specific null hypothesis, but rather to estimate a parameter or the strength of a relationship.

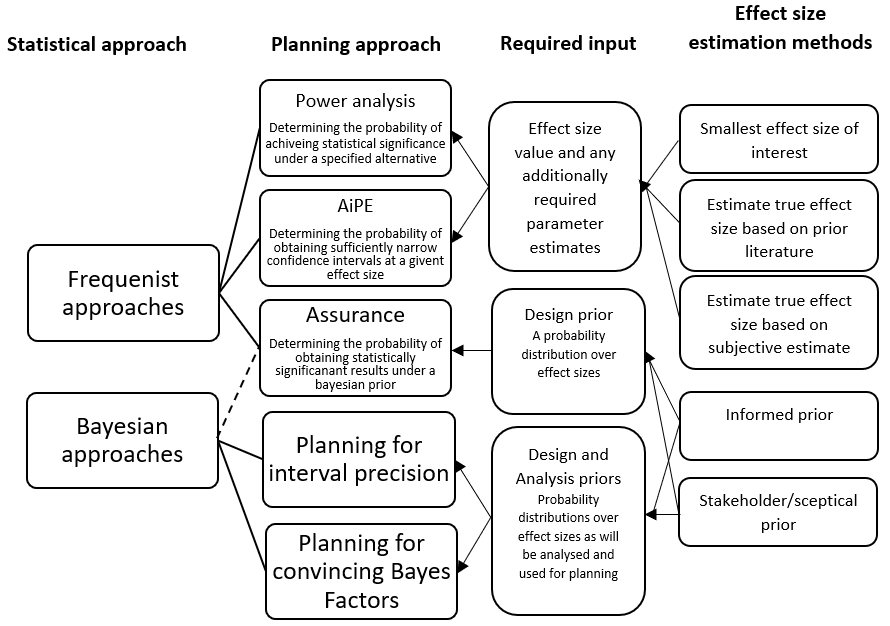


Figure 1. Illustration of the different approaches to effect size estimation determination in different statistical frameworks.

## Terms

Box 1

**Statistical power**

The probability of obtaining statistically significant results from a given statistical analysis under a specific alternative hypothesis.

**Power analysis**

The process of estimating the statistical power of a particular analysis given a set of specified population parameters and test characteristics (e.g., a sample size, effect size, alpha level, and the number of groups included)

**Accuracy in Parameter Estimation (AiPE)**

The process of estimating the probability of obtaining sufficiently precise confidence intervals under a set of specified population parameters and test characteristics (e.g., a sample size, effect size, confidence level, and the number of groups included)

**Assurance**

The probability that a frequentist statistical procedure will achieve a given goal (e.g., statistical significance or sufficiently precise confidence intervals) given a set of specified test and design characteristics (e.g., sample size, alpha level, the number of groups included in the analysis) and a Bayesian design prior over the effect size of interest.

**Sample size planning for Bayesian interval precision**

Estimating the sample size necessary to obtain a specified probability of obtaining sufficiently precise Bayesian credible intervals given a specified Bayesian Design and Analysis prior.

**Sample size planning for convincing Bayes Factors**

Estimating the sample size required to obtain a specific probability of obtaining sufficiently convincing Bayes Factors using a particular Bayesian analysis given a specified Bayesian Design and Analysis prior.

**Bayesian Design Prior**

A specified distribution over effect sizes to be used in the planning of studies (often a distribution of expected effect sizes)

**Bayesian Analysis Prior**

A specified distribution over effect sizes to be used in the planning of studies (often representing a sceptical observer or one of a number of default priors)

**Confidence level**

The confidence level of a confidence interval describes the probability that a particular statistical procedure will contain the true population parameter given that the statistical procedure’s assumptions are met.

## Approaches to formal sample size planning

In frequentist sample size determination, such as power analysis or Accuracy in Parameter Estimation (AiPE), researchers must specify an alternative hypothesis and research design in sufficient detail to determine the sampling distribution of the test statistic under a specific alternative hypothesis. This allows for an examination of the expected behaviour of a particular statistical analysis under the specified alternative hypothesis. For relatively simple designs (e.g., for a comparison of the mean scores of two independent groups or correlational analysis) the specification of a single standardised effect size (e.g., a Cohen’s *d* or Pearson’s *r*) and sample size characterises the sampling distribution under the alternative hypothesis adequately for power analysis (Cohen, 1988). For more complex designs (e.g., when covariates are to be included or when repeated measures designs are used) additional parameters may need to be specified such as the correlation between repeated measures and the number of included variables.

In the frequentist tradition, there is another approach to sample size determination typically called the Accuracy in Parameter Estimation (AiPE) approach (Maxwell, Kelley, & Rausch, 2007). In the AiPE approach, a researcher plans their sample sizes in order to obtain a given probability of obtaining sufficiently precise (i.e., narrow) confidence intervals under a specified alternative hypothesis and approach to statistical analysis. If researchers’ research goals are to estimate parameters with a given precision, as opposed to rejecting a null hypothesis, it makes sense that they would use this approach as opposed to statistical power analysis.

Whatever approach to statistical analysis is taken, when formal sample size planning is used it is important to understand how the method used to develop the alternative hypothesis (i.e., the effect sizes and parameter estimates used in power analysis) impacts the meaning of the results. Under no approach is it possible to discover the ‘true’ statistical power (or probability of obtaining narrow intervals) of a proposed analysis, as the true effect size is never known (Morey & Mayo, 2017, July 26th).

**What approaches are currently being used?**

In order to get an initial estimate of the research planning practices common in psychology I assessed the 121 empirical research articles published in the November 2017 to August 2018 issues of Psychological Science. Of the 121 empirical research articles published during this period 51 articles reported a power analysis, 42% of sampled articles (95% Wilson score interval [34%, 51%]). None reported using any other technique (e.g., AIPE to plan for precise interval estimates or Bayesian sample size planning methods). Of the reported power analyses, the most common approach was to effect size selection was to use a single previous study as the effect size, with 12 articles (10% of examined articles) reporting having done so. Despite the fact that pilot studies are (almost by definition) too small to reliably estimate the true population parameter value of interest, 3 studies (2% of articles) reported having estimated the effect size with this value. Almost as many used benchmarks from Cohen (1988; n = 9, 7% of articles). Six articles (5%) reported a sensitivity analysis, showing the effect size that the sample size gave them 80% power to detect. Seven articles (6%), did not provide any justification for the effect size they reported having used in power analysis, and 4 articles (3%) did not state the effect size that they used in a reported power analysis. Just 3 articles, 2% of those examined, reported that they adjusted their estimates for publication bias, and all of these articles used ad-hoc methods such as doubling the sample size that resulted from a power analysis or using the lowest reported effect for an intervention as opposed to the more sophisticated methods that have been proposed (e.g., Anderson et al., 2017; Perugini, Gallucci, & Costantini, 2014). See table [Psych sci] for the number and percentage of papers reporting each type of justification for the effect sizes reported in their power analysis along with multinomial confidence intervals on the percentages of papers in each group. This data is available from <https://osf.io/bmv2d/>.

Table [1]. *The number and percentage of papers reporting each type of justification for the effect sizes reported in their power analysis along with multinomial confidence intervals on the percentages of papers in each group.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  | 95% CI | |
| Effect size selection method | n | Percentage | LB | UB |
| No power analysis reported | 70 | 58 | 50 | 67 |
| Single previous study | 12 | 10 | 2 | 19 |
| Informal assertion of effect size | 7 | 6 | 0 | 15 |
| Medium effect benchmark from Cohen | 6 | 5 | 0 | 14 |
| Sensitivity analysis | 6 | 5 | 0 | 14 |
| No effect size stated | 4 | 3 | 0 | 12 |
| Average effect size in a set of studies (not a formal meta-analysis) | 3 | 2 | 0 | 11 |
| Effect size from a pilot study | 3 | 2 | 0 | 11 |
| Small effect benchmark from Cohen | 3 | 2 | 0 | 11 |
| Average effect size in a set of studies (not a formal meta-analysis), reduced for publication bias | 2 | 2 | 0 | 11 |
| Effect size from meta-analysis | 2 | 2 | 0 | 11 |
| Lowest effect size reported in a previous paper on this topic | 1 | 1 | 0 | 10 |
| Rule of thumb supported by power analysis | 1 | 1 | 0 | 10 |
| Smallest effect size from set of pilots | 1 | 1 | 0 | 10 |

Because of the infrequency with which the accuracy in parameter estimation (AIPE) approach (sample size planning to constrain confidence interval width) is reported (e.g., never in this relatively small sample of research), we focus on power analysis in this paper. However, when researchers are hoping to precisely estimate effects as opposed to investigating their presence, AIPE techniques provide a more appropriate method of planning sample sizes than power analysis (Maxwell, Kelley, & Rausch, 2008). Although the following discussion of effect size selection methods focuses on their implications for statistical power, these concepts apply equally to sample size planning for reliably narrow confidence interval width.

### Approach 1: estimating the population effect size

The most common approach to effect size selection for formal sample size planning in psychology is to develop an estimate of the effect size under study, an approach often presented as the only way of performing formal sample size planning (e.g., Kadam & Bhalerao, 2010; Kim & Seo, 2013). In this approach the researcher proposes a hypothetical scenario and says that if the true effect size was equal to their estimate (and all other statistical assumptions are met), the study would reach statistical significance on a given percentage of experiments (a value commonly and arbitrarily set at 80%). Often the estimated effect size is justified using a previous effect size from the literature or a meta-analytic effect size estimate. Failing that, the researcher uses their intuition to estimate the effect size. In so far as articles explicitly identified having used any of the available approaches to selecting effect sizes for power analysis in the examined issues of Psychological Science, almost all of the articles reporting a power analysis at least implicitly suggested that this was the goal of their power analysis (i.e., those using effect sizes from a single precious study, meta-analysis, average effects seen in previous research, effect sizes from a pilot study and informal assertions of the effect size).

While this approach is the most common, it may be the most error prone as minor differences in the choice of effect sizes will lead to large differences in the sample sizes that result (Wagenmakers et al., 2015). This is the only method of sample size determination that involves estimating the true effect size of the planned study, and the ‘required’ sample size is only as reliable as the effect size estimate that is used. If a researcher chooses arbitrarily, or adjusts their effect size estimate in order to achieve a certain level of apparent power at an achievable sample size, formal sample size planning is of little use. Any sample size greater than one has 80% power to detect a large enough effect size, while simultaneously being severely ‘underpowered’ to detect a sufficiently small effect size. The main difficulties in this approach are, firstly, selecting a sufficiently similar piece of previous research, and secondly in accounting for sampling variability and the impact of publication bias when using effect sizes from previous research.

#### Selecting a similar body of research

Given that as little as 1% of published psychology research psychology are direct replications of previous experiments (Makel, Plucker, & Hegarty, 2012) in the great majority of research psychologists have to base their effect size estimates on bodies of research that are markedly different from previous studies in at least some aspects. Even in so called direct replications, where the goal is to perform an experiment as closely as possible to an original study, it seems likely that there will be some small deviations from the original protocol that will mean that the assumption of both studies having a strictly identical population effect size may not be reasonable. This means that a researcher must use their judgement to identify a set of previous studies that are sufficiently similar to provide an estimate of the size of the effect under study, and must guess at the size and direction of any effect size differences from this pervious body of research, something that is an inherently subjective decision.

Although it is not uncommon for people to suggest that power analyses could be based on pilot studies (e.g., Kim & Seo, 2013), in so far as most pilot studies are not large enough to develop sufficiently precise estimate effect sizes for inferences about even the presence or absence of effects, using a point estimate from a pilot is likely to be of very little use (Albers & Lakens, 2018). However, it may be possible to use conservative estimates from aspects of the pilot study to inform aspects of sample size planning (e.g., using the value within a 95% CI around a standard deviation that leads to the largest required sample size) (Lancaster, Dodd, & Williamson, 2004). This same approach, taking the most conservative estimate from a pilot study’s 95% confidence interval, could be used for point estimates of the main statistic under study. However, it is likely that the resulting sample sizes will either be impractically large or the bounds will include 0.

Once a sufficiently similar body of research is identified, the parameters required for power analysis must be extracted. In the case of simple research designs this is usually possible from just the test statistics or reported effect sizes. Even when previous studies have not reported effect sizes directly many effect sizes (such as *d*, partial eta squared and *r*) that are required in popular power analysis computer programs (Faul, Erdfelder, Lang, & Buchner, 2007) are readily calculable from reported sample statistics and their associated degrees of freedom, or from reported means and standard deviations or standard errors. See chapter [effect size measures] for definitions and plain language descriptions of the most commonly used effect sizes for power analysis, and that chapters’ supplementary materials for a description of how they can be extracted from more commonly reported test statistics and effect sizes.

#### Accounting for sampling variability

If using this approach, once a sufficiently similar body of research has been identified, it is important to account the imprecision of previous effect size estimates and publication bias (Anderson & Maxwell, 2017). In the assessed Psychological Science articles, 12% reported directly using a power analysis based on a single effect size estimate from a previous study. Using effect sizes directly from the previous literature will lead to underestimates of the required sample size on 50% of occasions simply on the basis of sampling variability (Taylor & Muller, 1996). More seriously, because power is a concave function of sample size, using point estimates from previous studies to estimate required sample sizes will lead to mean levels of power being below the goal levels (McShane, Böckenholt, & Hansen, 2016). In other words, you get more severely punished for underestimating the effect size than you get rewarded for overestimating the effect size, which means that the mean power of a set of studies using point estimates from previous studies (or pilot studies) will be lower than their goal level of power.

Various methods have been developed to account for sampling variability in effect size estimates for power analysis (Anderson et al., 2017; McShane & Böckenholt, 2016; Perugini et al., 2014; and Taylor & Muller, 1996). Perugini et al. (2014) propose the “safeguard” approach to effect size selection for power analysis. This approach uses the lower bound of a confidence interval as opposed to the point estimate produced by an original study in order to have confidence (in the statistical sense of occurring at a particular long run relative frequency) that the study will have at least the chosen level of power to detect the unknown true population parameter. However, the actual confidence level achieved is only indicative of the true level of (statistical) confidence a researcher should have for direct replications in the absence of any type of selective reporting, an unlikely assumption in practice (Banks, Rogelberg, Woznyj, Landis, & Rupp, 2016). Even when all assumptions hold this method will tend to overestimate the required sample size on average (Perugini et al., 2014).

In order to deal with the fact that Perugini and colleagues (2016) method tends to overestimate the required sample size, McShane and Böckenholt (2016) propose the Power Calibrated Effect Size (PCES) approach to dealing with sampling variability. PCES takes a point estimate of the effect size and some measure of its uncertainty (e.g., a confidence interval or standard error) and uses this information to estimate the sample size required to ensure that the expected power (averaged over the likely distribution of effect sizes) is equal to a chosen level. In other words, this method aims to ensure that the mean level of power of studies choosing their sample size using this method will be at the chosen level, and not below that level as would be the case if people simply use reported point estimates of the effect size.

McShane and Böckenholt (2016) provide easily implemented formulas for tests of differences between independent and dependent means and proportions as well as correlation coefficients and host a website where these methods can be easily used (<https://blakemcshane.shinyapps.io/pces/>). McShane and Böckenholt (2016)’s methods for calculating PCES for these common statistical tests are equivalent to using the assurance approach (detailed below), using a prior distribution over the parameter of interest (e.g., a mean difference or Fisher Z transformed correlation coefficient) assumed to be normally distributed, centred on the point estimate of the parameter from a previous study, and to have a standard deviation equal to the standard error of the point estimate.

#### Accounting for publication bias

Both McShane and Böckenholt (2016) and Perugini et al. (2014)’s approaches assume that there is no publication or reporting bias in the literature. However, there is substantial evidence for publication bias in the behavioural sciences research literature (Johnson, Payne, Wang, Asher, & Mandal, 2017; Nelson, Simmons, & Simonsohn, 2018). For example, our recent examination of the effect sizes in replication studies suggests that, on average, the effect sizes seen in direct replications are on average as much as 20% smaller than those seen in original studies (95% credible interval [11%, 28%]) even when we condition on the true effect size being non-zero [effect size chapter]. If the reported point estimates of effect sizes seen in the psychological literature are likely to be overestimates of true effects planning sample sizes for replication studies without accounting for this will lead to overestimates of the unknown true power of replication studies (Anderson et al., 2017).

A method developed by Taylor and Muller (1996) and discussed in depth in Anderson et al. (2017) has been developed to account for publication bias. This method does not use the observed effect size from a previous study, but instead to uses a maximum-likelihood estimate of the effect size assuming a truncated test statistic distribution (Taylor and Muller suggest that it should be truncated at the critical value, i.e., at the test statistic which leads to statistical significance). This means that a study powered using this method is powered to detect the observed effect size calculated assuming that only the statistically significant test statistics are available or are considered for replication. Taylor and Muller (1996) suggest using the lower limit of a 95% confidence interval as a conservative heuristic. Mirroring Perugini et al. (2014)’s approach, using the lower bound of an a confidence interval leads to confidence that the study will have at least the goal level of power (i.e., the expected long run frequency of studies which will reach the achieved sample size) assuming that test statistics are only reported given that they reach statistical significance (and assuming that the true effect size under study is identical in the replication study as in the original study). Anderson and Maxwell (2017) provide an easy to use website that makes the implementation of their method much less complex than it may sound (see www.DesigningExperiments.com).

In suggesting that these methods of accounting for publication bias and sampling variability be used, it is worth noting that they are rarely used. Only three articles examined from Psychological Science reported attempting to account for publication bias. All three used ad hoc methods, one doubling the sample size they recruited from that returned by a power analysis that used effect sizes directly extracted from the literature, one reducing effect sizes by an arbitrarily chosen amount and one using the lowest effect size reported in a set of studies on their effect of interest. In fact, it seems entirely possible that these more complex methods of effect size adjustment may *never* have been used to plan sample sizes for research. A literature review performed by Anderson and Maxwell (2017) of articles citing Taylor and Muller (1996) and Perugini, Gallucci, and Constantini (2014) showed that none of the examined articles had actually used either of these methods to plan their sample sizes. This may be in part due to the scarcity of direct replications in the literature, the lack of awareness of these methods, or possibly the additional difficulty of using these methods. In so far as the technical difficulty of implementing these methods may disincentives their use, it may also be worth pointing out the simple heuristic suggested in Camerer et al. (2018). Camerer et al. (2018) suggest simply halving effect size estimates from the published literature, which they argue would capture the great majority of direct replication studies in which the effect size is non-zero.

### Approach 2: Smallest Effect Size of Interest

“Ideally, a study should be large enough to have a high probability (power) of detecting as statistically significant a clinically important difference of a given size if such a difference exists” CONSORT statement on selecting sample sizes for research (Moher et al., 2010, p. e10).

A more conservative approach to effect size selection than attempting to estimate the “true” effect size under study is to use sample size estimates based not on an estimate of the true effect, but rather to use the minimum effect size of theoretical, practical or clinical significance (sometimes called the Smallest Effect Size Of Interest, SESOI). If used in power analyses this approach means that if an effect as or larger than the smallest effect of interest is present it will be detected at a chosen probability or higher. Arguably, this is the most internally coherent and reliable method of selecting a sample size, as unlike the above method, it does not rely on estimating an unknown parameter, but rather in transparently making a subjective decision about the effect size that is minimally clinically or substantively important. None of the articles examined in Psychological Science reported having used this approach to plan the sample sizes they used.

#### Selecting a Smallest Effect Size of Interest

The selection of the smallest effect size of interest has recently been discussed in the psychology literature in discussions around equivalence testing (i.e., statistical testing to investigate whether an effect is significantly smaller than the SESOI; see Lakens, 2017; or Lakens, Scheel, & Isager, 2018 for an extended discussion of the process of selecting a SESOI). Briefly, the smallest effect of interest can be specified in terms of unstandardized units (i.e., in raw scores on a given measure) or standardised units (e.g., in Cohen’s *d* or ), depending on what can meaningfully be justified and the researcher’s background knowledge.

The SESOI can be justified on purely subjective grounds (e.g., by a researcher deciding that they do no feel an effect of 5 points on some scale or a Cohen’s *d* of .1 is not interesting to them), or in more objective terms (Lakens et al., 2018). Selection of the SESOI based on ‘objective’ standards has been most discussed in the clinical medicine literature. In medicine, finding the SESOI typically has two goals, one is in sample size planning, and the other is in determining what level of improvement would suggest that a change in treatment is advisable (Jaeschke, Singer, & Guyatt, 1989). Towards these ends, several methods of justification are commonly used. One method of SESOI justification from the medical literature is the Minimal Detectable Difference (MDD), equivalent to the idea of detection thresholds in psychophysics (Norman, Sloan, & Wyrwich, 2003). The MDD acts as a reasonable SESOI in cases where scale score differences that cannot be consciously detected can reasonably be said to not be important. This may often be the case in clinical research, where a treatment that does not cause a perceptible decrease in symptoms may be unlikely to be worth prescribing (Jaeschke et al., 1989). Other methods of determining the SESOI in the clinical literature that have been suggested include examining population level differences (Norman et al., 2003), and attempting to identify cut scores that can reliably predict other objective consequences (e.g., rehospitalisation). In non-clinical areas of the behavioural sciences, additional ‘objective’ justifications could be found in theoretical model predictions (e.g., when theories are well developed enough to make point or interval predictions), or by external constraints (e.g., if an educational intervention will only be implemented if it raises test scores by 10%).

#### Issues with the SESOI approach to sample size planning

Ensuring that a study has adequate sample sizes to reliably detect a minimum effect of interest can be particularly problematic in psychology research for several reasons. Deciding on a minimum effect size can be extremely difficult, especially in non-applied or basic research where any non-zero effect may be theoretically interesting. The canonical example of an area when any effect would be of interest is extrasensory perception, where any non-zero effect would radically change the way we understand the world. Less farcical examples exist where extremely small effect might be of interest, for example in studying health or environmental behaviour change interventions, where small effects on an individual level may have large economic, societal or environmental impacts. Because the sample size required to reach a given level of power can be made arbitrarily large as the effect size of interest is decreased (Neyman & Pearson, 1933), if the effect size of interest is “any non-zero effect” it will be impossible for a study to appear to be adequately powered.

These issues mean that although this may be the most conservative approach, its universal application could lead to counterproductive outcomes. Some of the most important research would seem prohibitively expensive if this approach was employed without regard to context, as the smaller the effect size of interest, the larger and more expensive performing “adequately powered” (according to the SESOI approach) studies would be. This seems likely to be the case for interventions where outcomes of interest may be low frequency but high impact (e.g., mortality or severe mental illness) and areas of research where research may have large and important societal outcomes. In these situations, if there is reason to think that larger effect should be expected, ensuring that a study is adequately powered to detect a minimum interesting effect will be an extremely inefficient use of resources. However, in cases where the marginal cost of additional participants is low and a reasonable SOSOI can be determined, this approach provides guarantees that the experiment is likely to provide meaningful evidence or adequate precision at a known rate given that an effect at least as large as the SESOI is present.

### Approach 3: Bayesian prior distribution

The final approach to effect size selection is to not specify a single alternative hypothesis, as is the case in the SESoI and effect size estimation approaches, but rather to use a Bayesian prior distribution over effect sizes, and to plan experiments to ensure that the outcome of interest is reached at some specified probability. If a Bayesian prior distribution is used to plan for frequentist statistical methods, this approach is often called planning for “assurance” (O'Hagan, Stevens, & Campbell, 2005). In the assurance approach, the prior distribution can represent the subjective probability of a researcher, that of a stakeholder or funder, or even a sceptical straw-man (Chen, Fraser, & Cuddeback, 2018). In the assurance approach, the prior distribution does not represent the expected effects under the alternative hypothesis, but rather a probability distribution over possible effect sizes, including the assigned probability of the null being true. The assurance no longer provides an estimate of the probability of obtaining statistical significance under the alternative hypothesis (or of obtaining sufficiently precise confidence intervals as in the case of AIPE), but instead estimates the probability of a researcher’s goals (e.g., statistical significance or sufficiently narrow CIs) being met given the specified prior distribution. The calculation of assurance under a specified prior distribution can be computationally complex, but reasonably easily implementable tools and methods have been developed (see Beavers & Stamey, 2018). This approach can also be used in planning experiments for more complex outcomes (e.g., “statistical significance given an effect of greater than a smallest effect size of interest or a statistically significant equivalence test if a smaller effect is present”), something that may be particularly appropriate when proposing expensive intervention or experiment.

#### Bayesian sample size planning

If researchers are planning to perform Bayesian statistical analysis, there are a number of different approaches to sample size planning. One method, developed by Schönbrodt and Wagenmakers (2017) is to plan a sample size based on the probability of obtaining sufficiently compelling and accurate Bayes factors, or Bayes Factor Design Analysis. In their approach a researcher must specify a design prior (distribution of expected effect sizes) under the null and alternative hypotheses. The researcher must also specify an analysis prior (a prior which will be used in the statistical analysis, e.g., a prior distribution designed to convince a sceptical audience). Random samples are then simulated from the Design priors for the null and alternative hypotheses, and Bayes Factors are computed using the analysis prior (Schönbrodt and Wagenmakers suggest repeating this “say, 10,000 times”). A researcher can then select an appropriate design based on an assessment how often a design provides compelling evidence (e.g., BF10 or BF01 > 10) under the null and alternative hypotheses at different sample sizes, and choose a sample size that maximises classification accuracy (Schönbrodt & Wagenmakers, 2017).

Another approach to Bayesian sample size determination is to specify a probability distribution over possible parameter values from the posterior distribution of an analysis of previous real data. The analyst can then sample parameter values (e.g., means and SDs) from the parameter value distribution, generate a set of simulated data, perform their statistical test on the simulated data, and check to see whether a particular goal condition has been met (e.g., sufficiently precise estimates, a sufficiently high or low Bayes factor, etc) (Kruschke & Liddell, 2017). The Bayesian and assurance approachs to sample size planning have the added benefit of acknowledging the uncertainty in parameter estimates, as opposed to the purely frequentist methods which tend to ignore this issue.

#### Issues with the Bayesian and assurance approaches to sample size determination

These tools are more flexible than their frequentist equivalents, and may be essential for planning research, budgeting, and in writing grant proposals when Bayesian analyses will be used for data-analysis. However, all of these methods, including the quasi-Bayesian assurance approach, require the specification of at least a design prior, the development of which is often a difficult task, and all currently require a level of technical expertise that is greater than that required in frequentist sample size planning where point-and-click interfaces are available. However, a number of tools have been already developed to enable researchers to develop reasonable prior distributions (e.g., Morris, Oakley, & Crowe, 2014), and it seems likely that more user friendly programs for Bayesian sample size planning (and analysis) will be developed as this approach to statistical analysis becomes more common (van de Schoot, Winter, Ryan, Zondervan-Zwijnenburg, & Depaoli, 2017).

### Conclusion

This paper is intended to act as an introduction to the various methods of selecting effect sizes for use in formal sample size planning, as well as the available methods of adjusting effect sizes for sampling variability and publication bias. Given that the majority of published papers used methods that imply they were aiming to estimate the true power of studies, and the infrequency with which articles attempted to deal with publication bias in deriving effect sizes, it is clear that these methods need to be more widely discussed. Researchers should be aware of the breath of methods available for sample size planning and the tools which have been developed to adjust effects for sampling variability and/or publication bias (see Anderson et al., 2017; McShane & Böckenholt, 2016; Perugini et al., 2014; and Taylor & Muller, 1996). It is also worth reinforcing the fact that although every study examined here that justified their sample size using a formal sample size planning tool used a power analysis, when researchers are primarily interested in precisely estimating a relationship it does not make sense to use a power analysis. In these cases, planning studies for estimate precision (or the equivalent in a Bayesian framework), would make more sense. See Maxwell et al. (2008) for a readable introduction to this approach in the frequentist framework.

The most appropriate type of sample size planning will differ according to the desires of the researcher and the type of analysis to be performed. In cases where it is particularly difficult to specify an effect size on a priori grounds, a reasonable approach may be to figure out the maximum sample size that can be recruited due to practical constraints such as funding or available time, and use this value to perform a sensitivity analysis, estimating the effect size that can be detected at a goal level of statistical power or with adequate precision. Equivalently, the power curve of an analysis (the power of the test over a range of possible population effect sizes given the maximum sample size achievable) could be examined in order for the researcher to understand the effect of different effect sizes on the statistical power of an analysis. If the effect sizes required to achieve adequate power or precision are larger than a researcher believes are likely or possible at the maximum sample size that is recruitable, it may not be advisable to not perform an experiment. These decisions are by their nature subjective, and the goal of formal sample size planning should be to ensure that researchers or funders have enough information to make informed decisions about whether and how to run experiments, not necessarily to attempt to select the “correct” sample size.

Finally, researchers may justifiably decide to perform research that is likely to have less that typical goal levels of power or which is likely to produce imprecise estimates of an effect. However, in these cases it becomes acutely important to ensure that the data will be available to future meta-analysts regardless of the statistical significance of results. Access to non-significant results is essential for the accurate estimation of effect sizes though meta-analysis (Anderson et al., 2017). Thankfully, newly developed tools such as pre-print servers like psyarxiv.com or data-repositories like figshare (figshare.com) and the Open Science Framework (osf.io), are now available. These services mean it is now possible for researchers to make your data and the results of analyses both available and discoverable outside of the traditional publication system.

**References**

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. doi:<https://doi.org/10.1016/j.jesp.2017.09.004>

Anderson, S. F., Kelley, K., & Maxwell, S. E. (2017). Sample-Size Planning for More Accurate Statistical Power: A Method Adjusting Sample Effect Sizes for Publication Bias and Uncertainty. *Psychological Science, 28*(11), 1547-1562. doi:10.1177/0956797617723724

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, 52*(3), 305-324. doi:10.1080/00273171.2017.1289361

Banks, G. C., Rogelberg, S. G., Woznyj, H. M., Landis, R. S., & Rupp, D. E. (2016). Editorial: Evidence on Questionable Research Practices: The Good, the Bad, and the Ugly. *Journal of Business and Psychology, 31*(3), 323-338. doi:10.1007/s10869-016-9456-7

Beavers, D. P., & Stamey, J. D. (2018). Bayesian sample size determination for cost-effectiveness studies with censored data. *PLOS ONE, 13*(1), e0190422. doi:10.1371/journal.pone.0190422

Biau, D. J., Kernéis, S., & Porcher, R. (2008). Statistics in brief: The importance of sample size in the planning and interpretation of medical research. *Clinical Orthopaedics and Related Research, 466*(9), 2282-2288. doi:10.1007/s11999-008-0346-9

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., . . . Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour, 2*(9), 637-644. doi:10.1038/s41562-018-0399-z

Chen, D.-G., Fraser, M. W., & Cuddeback, G. S. (2018). Assurance in Intervention Research: A Bayesian Perspective on Statistical Power. *Journal of the Society for Social Work and Research, 9*(1), 159-173. doi:10.1086/696239

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

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. doi:10.3758/bf03193146

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

Johnson, V. E., Payne, R. D., Wang, T., Asher, A., & Mandal, S. (2017). On the Reproducibility of Psychological Science. *Journal of the American Statistical Association, 112*(517), 1-10. doi:10.1080/01621459.2016.1240079

Kadam, P., & Bhalerao, S. (2010). Sample size calculation. *International Journal of Ayurveda Research, 1*(1), 55-57. doi:10.4103/0974-7788.59946

Kim, J., & Seo, B. S. (2013). How to Calculate Sample Size and Why. *Clinics in Orthopedic Surgery, 5*(3), 235-242. doi:10.4055/cios.2013.5.3.235

Kruschke, J. K., & Liddell, T. M. (2017). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*. doi:10.3758/s13423-016-1221-4

Lakens, D. (2017). Equivalence Tests. *Social Psychological and Personality Science, 8*(4), 355-362. doi:10.1177/1948550617697177

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. doi:10.1177/2515245918770963

Lancaster, G. A., Dodd, S., & Williamson, P. R. (2004). Design and analysis of pilot studies: recommendations for good practice. *Journal of Evaluation in Clinical Practice, 10*(2), 307-312. doi:10.1111/j..2002.384.doc.x

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in Psychology Research. *Perspectives on Psychological Science, 7*(6), 537-542. doi:10.1177/1745691612460688

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2007). Sample Size Planning for Statistical Power and Accuracy in Parameter Estimation. *Annual Review of Psychology, 59*(1), 537-563. doi:10.1146/annurev.psych.59.103006.093735

Maxwell, S. E., Kelley, K., & Rausch, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology, 59*(1), 537-563. doi:doi:10.1146/annurev.psych.59.103006.093735

McShane, B. B., & Böckenholt, U. (2016). Planning sample sizes when effect sizes are uncertain: The power-calibrated effect size approach. *Psychological Methods, 21*(1), 47-60. doi:10.1037/met0000036

McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016). Adjusting for Publication Bias in Meta-Analysis: An Evaluation of Selection Methods and Some Cautionary Notes. *Perspectives on Psychological Science, 11*(5), 730-749. doi:10.1177/1745691616662243

Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gøtzsche, P. C., Devereaux, P. J., . . . Altman, D. G. (2010). CONSORT 2010 Explanation and Elaboration: Updated guidelines for reporting parallel group randomised trials. *Journal of Clinical Epidemiology, 63*(8), e1-e37. doi:10.1016/j.jclinepi.2010.03.004

Morey, R. D., & Mayo, D. (2017, July 26th). A Poor Prognosis for the Diagnostic Screening Critique of Statistical Tests. Retrieved from osf.io/nepx9.

Morris, D. E., Oakley, J. E., & Crowe, J. A. (2014). A web-based tool for eliciting probability distributions from experts. *Environmental Modelling & Software, 52*, 1-4. doi:<https://doi.org/10.1016/j.envsoft.2013.10.010>

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's Renaissance. *Annual Review of Psychology, 69*(1), 511-534. doi:10.1146/annurev-psych-122216-011836

Neyman, J., & Pearson, E. S. (1933). The testing of statistical hypotheses in relation to probabilities a priori. *Mathematical Proceedings of the Cambridge Philosophical Society, 29*(4), 492-510. doi:10.1017/S030500410001152X

Norman, G. R., Sloan, J. A., & Wyrwich, K. W. (2003). Interpretation of Changes in Health-Related Quality of Life: The Remarkable Universality of Half a Standard Deviation. *Medical Care, 41*(5), 582-592. Retrieved from <http://www.jstor.org.ezp.lib.unimelb.edu.au/stable/3768017>

O'Hagan, A., Stevens, J. W., & Campbell, M. J. (2005). Assurance in clinical trial design. *Pharmaceutical Statistics, 4*(3), 187-201. doi:10.1002/pst.175

Perugini, M., Gallucci, M., & Costantini, G. (2014). Safeguard Power as a Protection Against Imprecise Power Estimates. *Perspectives on Psychological Science, 9*(3), 319-332. doi:10.1177/1745691614528519

Ren, S., & Oakley, J. E. (2014). Assurance calculations for planning clinical trials with time‐to‐event outcomes. *Statistics in Medicine, 33*(1), 31-45. doi:10.1002/sim.5916

Schönbrodt, F. D., & Wagenmakers, E.-J. (2017). Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review*. doi:10.3758/s13423-017-1230-y

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. doi:10.1080/03610929608831787

van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: The last 25 years. *Psychological Methods, 22*(2), 217-239. doi:10.1037/met0000100

Wagenmakers, E.-J., Verhagen, J., Ly, A., Bakker, M., Lee, M. D., Matzke, D., . . . Morey, R. D. (2015). A power fallacy. *Behavior Research Methods, 47*(4), 913-917. doi:10.3758/s13428-014-0517-4