# Chapter x.1: Approaches to formal sample size determination

Author: Felix Singleton Thorn

Researchers routinely have to decide upon the sample size they include in their research. If they decide to use formal sample size planning tools like power analysis to do so, they must specify the effect size and other parameters that are required for power analysis. There are two commonly used approaches reported by psychologists, estimating the expected effect size (Anderson, Kelley, & Maxwell, 2017), using the minimum interesting or clinically significant effect size (Biau, Kernéis, & Porcher, 2008). A second approach that is less commonly employed is to use a Bayesian prior distribution over effect sizes and parameters, either to estimate ‘assurance’, the probability that a study will be ‘successful’ based on an author’s outcome criteria (Ren & Oakley, 2014), or the probability of developing convincing evidence or precise interval estimates in Bayesian analyses.

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 effect, minimum effect of interest or using a Bayesian prior distribution may be appropriate in different circumstances, although they provide different information. This chapter outlines these different approaches to selecting effect sizes in formal power analysis and clearly explains the implications of each approach, as well as reinforcing the warnings that have been provided against using effect size estimates from pilot studies.

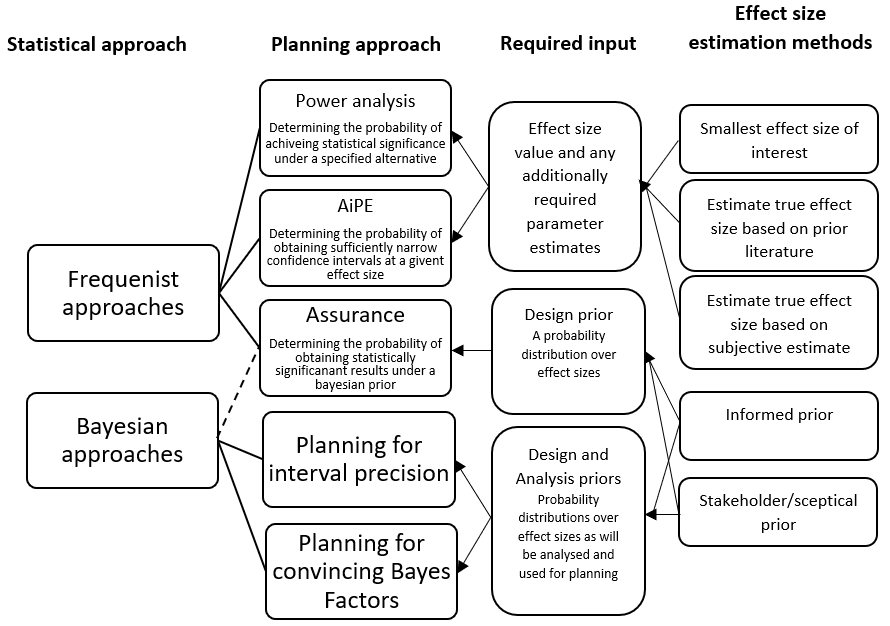


Figure [approaches to formal sample size planning]. Figure showing the different approaches to effect size estimation determination in different statistical frameworks.

## Terms

Box [Definitions]

**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.

Whatever statistical analysis approach is taken, 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) impacts the meaning of the results. Under no approach is it possible to discover the ‘true’ statistical power of a given statistical analysis, the true effect size is never known and these different approaches to estimating effect sizes for power analysis lead to very different interpretations of the estimated statistical power of analyses. It is arguable that there is no single true effect size on an a priori basis, as the small differences in variability in methods, situation and population sampled from will impact the unknown true effect size at least to some extent (Amrhein, Trafimow, & Greenland, 2018). The current paper focuses on the final level of Figure [approaches to formal sample size planning], on the different approaches to developing effect size estimates for use in power analysis, giving straightforward advice on how an interested researcher can meaningfully develop effect size estimates for use in a power analysis.

**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. The sample size included in this data-collection effort was determined in order to constrain multinomial confidence interval width to a maximum of 20%. 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%]). 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 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 reported having estimated the effect size with this value 2% of articles. Almost as many used benchmarks from Cohen (1988; n = 9, 7 % of articles). Other articles either reported a sensitivity analysis (showing the effect size that the sample size gave them 80% power to detect, n = 6 (5% of articles) in order to justify the obtained sample size. 7, 6% of articles, did not provide any justification for the effect size they reported having used in power analysis, and 4 articles (3% of all articles surveyed) did not state the effect size that they used in a reported power analysis. Just 3, articles, 2% of the examined articles, 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 [1] for the number and percentage of papers reporting each type of justification for the effect sizes reported in their power analysis, and for multinomial CIs around the percentage of papers reporting each type. (see <https://osf.io/bmv2d/> for the data behind the above description and 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.

Table [psych sci]. *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 | Percent | 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 |

### Method 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 true effect size, 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 x% (a value commonly and arbitrarily set at 80%) of occasions. 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 a single precious study, meta-analysis, average effects, effect sizes from a pilot study, 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). As this is the only method of sample size determination that involves estimating the true effect size of the planned study, 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, formal sample size planning is of little use. Any sample size greater than one has 80% power to detect a large enough effect size.

The main difficulties in this approach are, firstly, identifying a sufficiently similar body of similar research which from which adequately reliable estimates can be developed, and secondly accounting for simple random variability and the likely impact of publication bias on the observed magnitude of effects. 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 non-central aspects of the pilot study to inform aspects of sample size planning (e.g., using the upper bound of a 95% CI as an estimate of the standard deviation of a random variable) (Lancaster, Dodd, & Williamson, 2004).This same approach could be used for point estimates of the statistic under study, but it is likely that the resulting sample sizes will either be impracticably large or the bounds will include 0 suggesting that the sample size required would be approximately infinite.

#### Selecting a similar body of research

Selecting a sufficiently similar sample will often be an inherently subjective decision. 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 the effect size estimates on bodies of research that are markedly different in at least some aspects. The situation is simpler in the case of direct or partial replications, where a previous study provides a direct or reasonable estimate of the true effect size of the effect under study, although even then it may be appropriate to adjust for publication bias. A researcher must use their expert judgement to identify a set of previous studies which are sufficiently similar to act as reasonable estimates of the type of effect that could be reasonably expected from the type of research that is being performed. Once this body of research, parameters required for power analysis must be extracted. In the case of simple research designs, this is usually possible from just the test statistics and 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. See chapter [effect size measures] for definitions of the most commonly required effect sizes are, and of how they can be extracted from more commonly reported test statistics and effect sizes.

The degree to which selection of an adequately similar body of research is likely to be a problem is a complicated issue in that the more novel the experiment, the less indicative previous effects sizes may be of the current study’s likely effect, and the degree of difference that could be expected is a matter of judgement. Because of this fact, in the great majority of cases these types of power analyses act not as estimates of the true power of a study, but as a demonstration that the study was adequately powered to detect reasonably similar types of effects, thus likely able to reliably detect a reasonable person’s guess of the study’s effect if there was an effect. If the researcher conducting the experiment had a high degree of confidence in the estimated effect based on previous research, they would likely not perform the planned experiment.

#### Accounting for effect size imprecision and exaggeration

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). 12% of examined studies reported using a power analysis based on a single effect size estimate from a previous study. Even ignoring publication bias, imprecision in the parameter estimates chosen can be problematic. Using effect sizes directly from previous literature will lead to power estimates inflating the apparent power at least 50% of the time simply on the basis of sampling variability (Taylor & Muller, 1996). Because power is a concave function of sample size that asymptotes at 1, mean power from a set of studies using the point estimate from previous studies to power their experiments will on average lead to power being below desired level (Blakeley, Ulf, & Karsten, 2016).

Publication bias compounds this issue, as the point estimates of effect sizes seen in the psychological literature are likely to be overestimates of true effects (Anderson et al., 2017). Evidence from projects that have systematically attempted to reproduce bodies of psychology research suggest that the impact of publication and reporting biases may be to increase reported effects by as much as 30 – 50% (Camerer et al., 2018; Open Science Collaboration, 2015). It is difficult to accurately account for publication bias in meta-analysis, and it is likely that many of the effects sizes reported in meta-analysis are also upwardly biased (Ferguson & Brannick, 2012; Hartwig, Davey Smith, Schmidt, & Bowden, 2018; Thompson et al., 2011).

However, methods have been developed to account for sample variability inherent point estimates of effects sizes or parameters (see Anderson et al., 2017; McShane & Böckenholt, 2016; Perugini et al., 2014; and Taylor & Muller, 1996). These methods are all designed to estimate the sample size required to adequately power a direct (or exact) replication study. The simplest method of adjusting effect sizes, suggested in Camerer et al. (2018), is halving effect size estimates from the published literature to arrive at a reasonable estimate of the true effect size. Perugini et al. (2014) propose safeguard power, using the lower bound of a 100-x% confidence interval as opposed to the point estimate in order to have 100-(x/2)% confidence (in the statistical sense), that the study will have adequate power to detect the true population parameter. However, the actual confidence level achieved is only indicative of the true level of confidence a researcher could have in the absence of any type of selective reporting, an unlikely assumption in practice. Even assuming that all assumptions hold this method will tend to overestimate the required sample size on average (Perugini et al., 2014).

A third approach has been proposed McShane and Böckenholt (2016) to address this issue that the safeguard power approach tends to overestimate effect sizes; the Power Calibrated Effect Size approach (PCES). PCES produces an effect size estimate that accounts for the variability in the point estimate such that the expected power of a study using this effect size will be powered at a chosen level of power. PCES is equivalent to the assurance approach (detailed below), using a prior distribution over the parameter of interest assumed to be normally distributed with a mean set at the point estimate from previous studies and a standard deviation set to the standard error of the test statistic. McShane and Böckenholt (2016) provide easily implemented formulas for tests of differences between independent and dependent means and proportions as well as correlation coefficients. They also provided a website at which these methods can be used at <https://blakemcshane.shinyapps.io/pces/>.

A final 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 maximum-likelihood estimate of the effect size assuming an uncensored test statistic distribution (e.g., an F distribution), but instead to uses a maximum-likelihood estimate of the effect size assuming a truncated test statistic distribution (they suggest that it should be truncated at the critical value, i.e., at the test statistic which leads to statistical significance). The lower bound of an 100-x% confidence interval around the effect size this can be used as the effect size measure in order to achieve 100-(x/2)% confidence (again, in the statistical sense) assuming that F-statistics are only reported given that they reach statistical significance. Anderson and Maxwell (2017) do provide an easy to use website that makes the implementation of their method much less complex that it may sound (www.DesigningExperiments.com).

However, methods for adjusting for publication bias are rarely used. This means that that some the most common methods of effect size selection (using effects from previous studies, either individual studies or from collections of studies) are likely to be optimistic estimates of power of the proposed research. Only three articles examined from Psychological Science reported attempting to account for publication bias. All three used reasonably ad hoc methods, one doubling the sample size they recruited from that returned by a power analysis using effect sizes from the literature, one reducing effect sizes by an arbitrary amount and one using the lowest effect size reported in a set of studies on their effect of interest. Furthermore, it seems these more complex methods of effect size adjustment may have never been used. 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.

These methods of adjusting effect sizes rely on researchers being able to identify a sufficiently similar piece of research. It may not be uncommon for a researcher performing novel research to not feel capable of identifying a sufficiently similar set of studies for a power analysis to be meaningfully performed if the power analysis relies on a plausibly accurate effect size estimate to be generated. If a researcher does not feel capable of developing an accurate effect size estimate, the more conservative an estimate that used to ensure that a planned study will be able to reliably detect a lower bound estimate of the effect if it is in fact present.

### 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 cautious approach 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 (or the Smallest Effect Size of Interest, SESoI). This is the method of sample size determination that leads to the largest sample sizes being determined in the AiPE or power analysis frameworks. 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 unknowable 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 explicitly noted that they were planning for the smallest effect of interest.

#### 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 SOSOI based on ‘objective’ standards has been most discussed in the clinical medicine literature. In medicine, finding the SESOI has two ends, 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). 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 is often the case in clinical research, where a treatment that does not cause a perceptible decrease in symptoms is 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 psychology research, 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 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. 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 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), meaning that if the effect size of interest is “any non-zero effect” it will be impossible for a study to appear to be adequately powered to detect the minimum effect of interest. Furthermore, in situations where 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 extremely inefficient.

Although this may be the most conservative approach, its strict application would lead to some of the most important research seeming prohibitively expensive and never being performed. Given that areas of research where small effects are be of particular interest, for example in interventions where outcomes of interest may be low frequency but high impact (e.g., mortality or severe mental illness), a large proportion of these studies would appear to require such high sample sizes so as to be impossible to practically perform, despite the fact that reasonably well-powered experiments may be possible given the unknown true impact of a set of interventions. However, in cases where the marginal cost of additional participants is low and reasonable SOSOI can be determined, this approach provides assurance that the experiment is likely to provide meaningful evidence or adequate precision given that the minimum effect 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 Bayesian methods for sample size planning. If a Bayesian prior distribution is used to plan for frequentist statistical methods, this approach is often called “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 the researchers’ prior distribution including their assigned probability for the null being true. The goal of assurance, the value that this approach estimates, is different from the above approaches. Assurance no longer provides an estimate of the probability of obtaining statistical significance under the null 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, or even correct classification of the presence or absence of effects using more complex decision rules) 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 (Beavers & Stamey, 2018).

#### Bayesian sample size planning

If researchers are planning to perform Bayesian statistical analysis, Schönbrodt and Wagenmakers (2017) developed a method to plan for sample size planning for sufficiently compelling and accurate Bayes factors. In their approach a researcher must specify a design prior (distribution of expected effect sizes) under the null, as well as the 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 under the null and alternative hypotheses (Schönbrodt & Wagenmakers, 2017).

Another approach to Bayesian sample size determination is to specify a probability distribution over possible parameter values often generated from the posterior distribution from a previous analysis of real or idealised data. Using Markov chain Monte Carlo sampling, the analyst can then sample parameter values (e.g., means and SDs) from the parameter value distribution, generate a set of simulated data, test 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). Bayesian power analysis has the added benefit of accounting for uncertainty in parameter estimates, as opposed to the frequentist methods which tend to ignore this issue.

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

If sample size needs to occur for Bayesian analyses, either for the purposes of research planning, for grant proposals etc. these tools are essential, and are more flexible than the frequentist versions. However, they do currently require a level of technical expertise that is greater than that required in frequentist sample size planning where point and click tools have been developed for the most common versions. Eliciting prior distributions can in of itself be difficult, although a number of tools have been developed to enable researchers to develop reasonable prior distributions (e.g., Morris, Oakley, & Crowe, 2014).

### Conclusion

The most appropriate type of sample size planning will differ according to the desires of the researcher, the type of analysis to be performed, and the desired role of the researcher. It is clear that the great majority of reported power analyses (at least those recently published in Psychological Science) tend to use effect sizes that are fairly simply derived and may be likely to underestimate the power of published studies even on the basis of sampling variability alone, an issue that is compounded by publication bias inflating effect sizes (Perugini et al., 2014). Given that the great 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 researchers should at least be aware of the breath of methods available to account for effect size inflation, and keep in mind that this is not the only method of coherently planning experiments.

Although focus has often been placed on increasing the sample size of research to increase power (e.g., Cohen, 1962 which relagates mention of other methods of increasing power to a footnote), other methods of increasing the power of statistical tests exist. Reducing measurement error, error variance, using repeated measures designs, increasing the alpha level, or increasing the size of the effect will lead to higher effect sizes and commensurately power in most cases (Loken & Gelman, 2017; Müller & Szegedi, 2002). One reason these other methods are often ignored is that these other elements are often assumed to be fixed by custom, already be optimised for maximum power, or difficult for the researcher to alter (Cohen, 1962). The alpha level used (which also impacts statistical power) is almost never changed, and then almost only ever made more strict leading to lower power (Gigerenzer, 2004). The effect size tends to be considered fixed, or assumed to already be maximized (Lakens & Evers, 2014). Other parameters (e.g., correlations between repeated measurements) are often thought to be outside of a researcher’s control. However, the experimental design will often be alterable (i.e., it may be possible to use a repeated measures design as opposed to a between subjects design), and it may be possible to prioritize obtaining highly reliable measures of constructs under study in order to maximize effect sizes (by reducing measurement error) and increasing power.

Ideally, researchers using frequentist statistical procedures would base their power analyses on the minimum effect size of interest in order to maximise the probability that they will obtain sufficiently precise estimates or have high statistical power if an effect of the minimum size or greater is present. However, it may be difficult to either establish a minimum effect size of interest, or the resulting effect size may be implausibly high. In these cases, when estimating the effect of interest it is important to recognise that the effect sizes in the literature are likely to be inflated to some unknown amount, and conservative estimates can be used to help ensure that experiments are likely to include enough participants. This can be done using the formal methods proposed by citing Taylor and Muller (1996) and Perugini, Gallucci, and Constantini (2014) to estimate the effect sizes of interest, or alternatively using easier informal adjustments like proposed in Camerer et al. (2018).

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 you can recruit and use this value to perform a sensitivity analysis (varying any other parameters that must be set for a power analysis or placing them at conservative estimates), estimating the effect size that can be detected at a goal level of statistical power or with adequate precision. Equivalently, an examination of 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 range of possible effect sizes that are reliably detectable. If the effect sizes required to achieve adequate power or precision are larger than you believe are likely or possible at the maximum sample size that is recruitable, consider not performing the experiment. If the experiment is going to go on in any case, assuming that researchers have strong incentives to obtain statistically significant results, it becomes acutely important to take measures to protect yourself from biases that may increase your false positive rate and to ensure that future meta-analysts will have access to your results regardless of the statistical significance or size of the observed effects. These imprecise estimates or non-significant results are essential if scientific research is to be able to accurately estimate effect sizes of some intervention (Anderson et al., 2017). Tools like pre-registration help ensure that researchers can adequately distinguish between the pre-planned confirmatory analyses and any exploratory analysis, and posting the results on pre-print servers like psyarxiv.com ensure that the results are accessible to any future meta-analysists regardless of the statistical significance of results.

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>

Amrhein, V., Trafimow, D., & Greenland, S. (2018). Abandon statistical inference. *PeerJ Preprints, 6*, e26857v26851. doi:10.7287/peerj.preprints.26857v1

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

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

Blakeley, B. M., Ulf, B., & Karsten, T. H. (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

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. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal of Abnormal and Social Psychology, 65*(3), 145-153. doi:10.1037/h0045186

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

Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychol Methods, 17*(1), 120-128. doi:10.1037/a0024445

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics, 33*(5), 587-606. doi:<http://dx.doi.org/10.1016/j.socec.2004.09.033>

Hartwig, F. P., Davey Smith, G., Schmidt, A. F., & Bowden, J. (2018). The median and the mode as robust meta-analysis methods in the presence of small study effects. *bioRxiv*. Retrieved from <http://biorxiv.org/content/early/2018/03/26/288050.abstract>

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>

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., & Evers, E. R. K. (2014). Sailing From the Seas of Chaos Into the Corridor of Stability. *Perspectives on Psychological Science, 9*(3), 278-292. doi:10.1177/1745691614528520

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

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science, 355*(6325), 584. doi:10.1126/science.aal3618

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

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

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

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>

Müller, M. J., & Szegedi, A. (2002). Effects of Interrater Reliability of Psychopathologic Assessment on Power and Sample Size Calculations in Clinical Trials. *Journal of Clinical Psychopharmacology, 22*(3). Retrieved from <http://journals.lww.com/psychopharmacology/Fulltext/2002/06000/Effects_of_Interrater_Reliability_of.13.aspx>

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

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251). Retrieved from <http://science.sciencemag.org/content/349/6251/aac4716.abstract>

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

Thompson, S., Ekelund, U., Jebb, S., Lindroos, A. K., Mander, A., Sharp, S., . . . Wilks, D. (2011). A proposed method of bias adjustment for meta-analyses of published observational studies. *International Journal of Epidemiology, 40*(3), 765-777. doi:10.1093/ije/dyq248

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