# x.1 How to choose effect sizes for sample size determination

There are three main approaches to selecting effect sizes for use in formal sample size planning. One approach is to estimate the expected effect size (Anderson, Kelley, & Maxwell, 2017). Another approach is to use the minimum interesting or clinically significant effect size (Biau, Kernéis, & Porcher, 2008). A third approach is to use a Bayesian prior distribution over effect sizes and parameters in order to estimate ‘assurance’, the probability that a study will be ‘successful’ based on an author’s outcome criteria (Ren & Oakley, 2014). Any formal sample size determination posits a hypothetical scenario (or probability distribution over different scenarios in the case of Assurance) and is only meaningful with regards to the proposed parameter values or prior distribution. All three methods of effect size selection 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 selecting sample sizes based on each approach. One of the difficulties researchers have when performing formal sample size planning is appropriately selecting parameters for use in power analysis [cite interviews and survey]. This chapter fills a clear gap in the research literature by giving a clear and succinct explanation of what information is given under these different approaches to selecting effect sizes and where they will be appropriate.

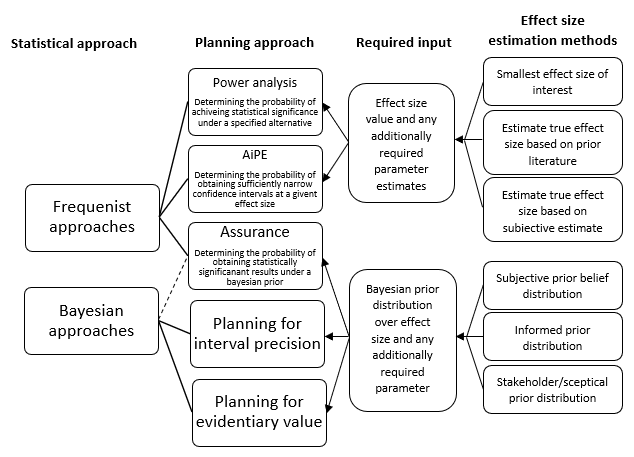


Figure [approaches to formal sample size planning]. Figure showing the different approaches to effect size estimation determination in different statistical frameworks.

## 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 in sufficient detail to determine the sampling distribution of the test statistic under the 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 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.

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. 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 and how an interested researcher can meaningfully develop effect size estimates for use in power analysis.

### 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% (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 this approach the required sample size is only as reliable as the effect size estimate that is used. 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 the presence or absence of effects using a point estimate from a pilot is likely to be of very little use (Albers & Lakens, 2018). 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, as any sample size greater than one has 80% power to detect a large enough effect size.

Although 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 size 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. The main difficulties in this approach are, firstly, identifying a sufficiently similar body of similar research which is reliable enough for estimates to be developed, and secondly accounting for simple random variability and the likely impact of publication bias on the observed magnitude of effects.

#### Selecting a similar body of research

The degree to which this 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, it seems fair to say that 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.

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 or partial 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. 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 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.

#### 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). Even ignoring publication bias imprecision in parameter estimates can be problematic, using effect sizes directly from previous literature will lead to power estimates inflating the apparent power approximately 50% of the time simply on the basis of sampling variability alone (Taylor & Muller, 1996). 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 2 fold (Open Science Collaboration, 2015) [add the Wagenmakers and Science/nature reproduction project paper here]. 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 publication bias and the sample variability inherent point estimates of effects sizes or parameters (see Anderson et al., 2017; McShane & Böckenholt, 2016; Perugini, Gallucci, & Costantini, 2014; and Taylor & Muller, 1996).

### 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).

It may not be uncommon for a researcher performing novel research to not to feel capable of identifying a sufficiently similar set of studies for a power analysis to be meaningfully performed. If a researcher does not feel capable of developing an accurate effect size estimate, a conservative estimate can be 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. A more extreme version of using a lower bound estimate to base their effect size estimate is a researcher basing their sample size not on an estimate of the true effect, but the minimum effect size of theoretical, practical or clinical significance (or the Smallest Effect Size of Interest, SESoI). In so far as people would not plan experiments if they expect effects that are smaller than the minimum interesting effect size, 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 theoretically meaningful 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.

**Issues with this approach**

However, ensuring that a study has adequate sample sizes to reliably detect a minimum effect of interest can be 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. However, in cases where the marginal cost of additional participants is low (e.g., in MTurk studies or when a study is brief and non-invasive and there is a readily available participant pool), this approach provides assurance that the experiment is likely to provide meaningful evidence or adequate precision given that the minimum effect is present.

Although this may be the most conservative and reliable approach, strict application of this approach would lead to some of the most important research seeming prohibitively expensive. 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.

### 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 provide a Bayesian-style prior distribution over effect sizes, either representing the subjective probability of a researcher, that of a stakeholder or funder, or even a sceptical straw-man (Chen, Fraser, & Cuddeback, 2018). If a single effect size is insufficient to adequately specify the alternative hypothesis, multiple prior distributions over each parameter must be determined (Chen et al., 2018). Specifying a prior is required for Bayesian statistical analyses, and using this approach aligns well with pre-registration of prior belief distributions, ensuring that belief priors are actually specified before analysis preventing them from being influenced by the data (Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012).

If a Bayesian prior distribution is used to plan for frequentist statistical methods, this approach is often called “assurance” (O'Hagan, Stevens, & Campbell, 2005). The goal of assurance, the value that this approach estimates, is different from the above approaches. Assurance no longer provides an estimate of the statistical power (or of the probability of obtaining sufficiently precise confidence intervals 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 a stated 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 most for cost effectiveness research in the context of for medical trials (Beavers & Stamey, 2018).

If a researcher has clear statistical outcomes in mind (e.g., that they achieve statistical significance or can support a null result via equivalence testing, etc.), they feel comfortable specifying a Bayesian prior distribution over effect sizes and other parameters, and they are interested in estimating the probability of those end goals being met this approach allows them to do so. These criteria may be rarely met.

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

Yin, K., Choudhary, P. K., Varghese, D., & Goodman, S. R. (2008). A Bayesian approach for sample size determination in method comparison studies. *Statistics in medicine*, *27*(13), 2273-2289.

O’Hagan, A., & Stevens, J. W. (2001). Bayesian assessment of sample size for clinical trials of cost-effectiveness. *Medical Decision Making*, *21*(3), 219-230.

AND MENTION THAT more complex versions which explicitly take into account the costs are also available in the medical literature, e.g.,

Further elaborations on these approaches, for example explicitly performing a cost benefit analysis during sample size planning, have been developed in the medical and pharmaceutical testing literature and tend to be labour intensive and rely on well developed expert judgements, and are not elaborated here, interested readers could see … for an in depth exploration of these approaches. All approaches may be reasonable in different circumstances, and both have their limitations. In all approaches to sample size selection, effect sizes or the prior distribution of effect sizes must be chosen appropriately for formal sample size planning to provide meaningful results.

|  |  |
| --- | --- |
| **Approach to effect size selection** | **Implication** |
| Smallest effect size of interest (SESoI) | Effect sizes as large or greater than the smallest effect size of interest have at least the chosen probability of being detected |
| Estimated true effect | Effect sizes as large or greater than the estimated size of interest have at least the chosen probability of being detected. Only as accurate as the effect selected. |
| Bayesian prior distribution (assurance) | Gives the probability of obtaining a set of statistical results (often statistical significance) given a Bayesian prior distribution over possible parameter values. Only as accurate as the prior distribution. |

**Current practices**

In order to get an initial sense of they types of practices common in psychology I assessed the July and August 2018 issues of Psychological Science. The sample is not representative of psychology articles in general, and the estimates are not precise enough to make any claims about the general trends in the examined journal, but nonetheless, the observed patterns are stark and align with my intuitions about research as a whole. Of the 30 empirical research articles published in this period 50% (n = 15) reported a power analysis. Of the reported power analyses, the most common approach (used in 9 of 15 cases) was to estimate the true effect of the intervention. Most of these articles used a point estimate from a single previous study (n = 6) to estimate the effect of an intervention. Just two articles reported using effect sizes from meta analyses toward the same goal, and one used the effect size seen in a pilot study. The other articles either reported a sensitivity analysis (showing the effect size that the sample size gave them 80% power to detect, n = 3) in order to justify the obtained sample size, or they used benchmarks from Cohen (1988; n = 2) or did not provide any justification for the effect size used in power analysis (n = 1), making it unclear whether their estimate was of the minimum effect of interest or an estimate of the true effect size of the intervention (see <https://osf.io/bmv2d/> for the data behind the above description).

There are several possible reasons for the scarcity of Assurance calculations in psychological research. Firstly, researchers appear to rarely use any of the more complex approaches to These methods are rarely 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.

**3.8 Increasing precision or power without increasing the sample size included in a study**

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 lowered 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 power.

This paper has not considered other approaches to sample size planning that do not require estimation of the effect size

There are other suggested approaches to sample size planning that do not

Equivalently, an examination of the power curve of an analysis (the power of the test over a range of possible population effect sizes) could be examined in order for the researcher to understand the range of possible effect sizes that are reliably detectable in order to decide whether it is advisable for an experiment to be conducted

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 or placing them at conservative estimates), estimating the effect size that can be detected at a goal level of statistical power. If the estimated effect size seems larger than you believe is likely, consider not performing the experiment. If the experiment is going to go on in any case, take care to ensure that the results will be available to meta-analysts and other researchers regardless of the statistical significance of results (e.g., by posting results on a publicly searchable archive like psyarxiv.com), and ensure that the analysis plan is pre-registered in order to allow yourself to distinguish between the pre-planned confirmatory analyses and any future exploratory analysis.

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

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

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>

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

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

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

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

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

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

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An Agenda for Purely Confirmatory Research. *Perspectives on Psychological Science, 7*(6), 632-638. doi:10.1177/1745691612463078