**x.1 Approaches to formal sample size determination**

In order to perform formal sample size determination, such as power analysis or Accuracy in Parameter Estimation methods, 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 analysis under the 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. One of the major difficulties often cited by researchers in performing a power analysis is that they have trouble developing appropriate parameters for use in power analysis [cite interviews and survey]. One possible reason for this issue is confusion over the appropriate way to select effect sizes for power analysis.

There are multiple recommended approaches to determining an effect size for power analysis. One approach is to use the minimum interesting or clinically significant effect (Biau, Kernéis, & Porcher, 2008), another is to estimate the expected effect size (Anderson, Kelley, & Maxwell, 2017). Both approaches may be reasonable in different circumstances, and both have their limitations. In both approaches to sample size selection, effect sizes must be chosen appropriately in order for power analysis to provide meaningful results. This paper outlines the different approaches to selecting effect sizes for use in power analysis and outlines how researchers can approach sample size planning with different goals in mind in order to ensure that their research is likely to accurately guide their inferences. Additionally, this paper discusses previously proposed benchmarks for effect sizes for use in power analysis, and provides a systematic review of previous studies which have developed empirical benchmarks from bodies of literature in order to show the distribution of effect sizes in different areas of research. Although in most scenarios standardised effect size benchmarks should not be used as the sole basis for a power analysis, in some cases little other guidance is available.

**What is formal sample size planning?**

There are several types of formal sample size planning, and the appropriate type of formal sample size planning depends on the inferential approach that is taken in the analysis of data. The most common type of formal sample size planning is power analysis, which allows researchers to determine the probability of achieving statistically significant with a particular statistical tests under a specified alternative hypothesis. Another type of formal sample size planning has been called the Accuracy in Parameter Estimation (AiPE) approach (Maxwell, Kelley, & Rausch, 2008), which refers to estimating the sample size requires in order to ensure that the confidence intervals that developed are sufficiently narrow. If Bayesian analyses are to be used, there are equivalent methods which have been developed and are under development which aim to ensure that that sufficiently precise estimates or sufficiently convincing evidence is likely to be developed (Kruschke & Liddell, 2017). One benefit of these Bayesian methods is that it is possible to distributions over parameter values which represent not just a single alternative hypothesis, but researchers’ beliefs or posterior distributions from a pilot or similar study.

Importantly, formal sample size sample size determination is only valid in regards to the particular statistical test that is to be used in analysis of the data. If an independent samples *t*-test is to be used to assess the outcome of a study, power analysis should be performed for an independent samples *t*-test. If Wilson score intervals are to be used to guide inferences, sample size planning should be taken to ensure that sufficiently narrow confidence intervals are developed to adequately guide inferences for the researcher’s purpures.

Even in cases where the upper limit of an achievable sample size is fixed (e.g., a budget is fixed), these same methods for determining sample sizes can be used to help to investigate whether it is worthwhile to perform an experiment. If a study is likely to be extremely underpowered or unlikely to provide sufficiently precise parameter estimates to guide inferences, the information gained through the experiment may not be worth the costs associated with performing the study. These tools ensure that researchers can be aware of this before they use their limited resources on studies which are unlikely to provide meaningful results regardless of the whether the effects are present or absent.

Whatever statistical analysis approach is taken, it is important to understand the likely behaviour of your statistical test under alternative hypotheses.

**Approach 1 – Minimal clinically significant difference**

In general, the smaller the expected effect size, the larger the width of confidence intervals and the lower the statistical power of research is expected to be (Cohen, 1992). Ensuring that adequate sample sizes are reached to have sufficient precision or adequate statistical power to detect a minimally interesting effect size ensures that if a larger effect is present even greater levels of power or precision will be reached. 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 of interest. Under this approach 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). This means that if the effect size of interest is “any non-zero effect” it is not possible for a study 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 or when a study is brief and non-invasive), this approach provides assurance that the experiment is likely to provide meaningful evidence or adequate precision given that the minimum effect is present.

**Selecting a minimum effect of interest**

Method 2 – estimating a likely effect size

On the other hand, using power analysis to develop a required sample size based on an estimated effect size is only as reliable as the effect size estimate that is used. If a researcher chooses arbitrarily or has an optimistic estimate of the true effect, a study may be underpowered to detect a more realistic effect size. Furthermore, if point estimates of effect sizes from the literature are used without accounting for random variability and possible impact of publication bias on the magnitude of effects, estimates are often extremely optimistic (Anderson et al., 2017). However, methods have been developed to ensure that reasonable effect size estimates can be developed accounting for publication bias and estimation uncertainty if adequate numbers of sufficiently similar previous studies are available (Anderson et al., 2017; McShane & Böckenholt, 2016).

It may not be uncommon for a researcher performing novel research to not to feel capable of developing an accurate effect size estimate, and in those cases 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. Alternatively, 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.

sample size planning should occur with the researcher’s specific outcome in mind. If a researcher is planning to base their inference on an effect size estimate and confidence intervals, sample size determination should aim to develop sufficiently narrow confidence intervals. If inferences are to be based on a traditional statistical test, power analysis should be used to plan for a sample size which provides a sufficiently high probability of the statistical test to be used returning a significant result at a given effect sizes.

DESIGN ANALYSIS HERE? Tie into earlier type s / type M error discussion

**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 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.

**Interpreting effect sizes**

Power analysis for a t-test for mean differences between independent groups can either take as input parameters the mean difference along with an estimate of the expected SD of the groups, or a direct estimate of the standardized effect size between groups. The most commonly referenced program for performing power analysis, g\*power [Honours research too, also see later chapter], accepts either as input, although the default input is Cohen’s *d* (Faul, Erdfelder, Lang, & Buchner, 2007)*.*

Proportion overlap (U) -



*Figure [Cohen’s d as population distributions]*. Population distributions with a mean difference of .2, .5, .8 and 1.2 Cohen’s d, along with the percentage overlap between populations (calculated assuming that populations are normally distributed, have equal variance, and equal sample sizes, using equations from (Reiser & Faraggi, 1999)).

**Effect size benchmarks**

The use of standardized effect size benchmarks such as those proposed in Cohen (1962) is often criticized on the grounds that these benchmarks are not empirically validated (i.e., the “medium” effect size benchmark is not the mean or median effect across psychology research), and that using standardized effect size benchmarks are likely to be poor estimates of the actual effect size of any particular experiment. Not only do they not adhere to the average size of effect seen in a particular research subfield, but researchers will often have additional information about the effect sizes that could plausibly be expected from a given experiment.

The “small” and “large” effect size benchmarks examined

To get an intuitive sense of what each effect size is, it is worth examining another effect size, the percentage point group overlap (always calculated assuming that the distribution of the two groups is normal PERCENTAGE OVERLAP FORMULA -> (Reiser & Faraggi, 1999).

In fact, the use of standardized effect sizes has been criticized in of itself, on the grounds that they make it more difficult to extract the main an (e.g., Cohen’s *d*) not only that, but ??

However these benchmarks are useful in the case of power surveys, in which case we are not interested in estimating the power of any individual experiment, but instead can examine the hypothetical power of a body of studies at a variety of plausible effect sizes.

Look at:

Interpreting the magnitudes of correlation coefficients. (Hemphill 2003)

**Benchmarks**

4.1.3 True average effect sizes seen in areas of psychology research

Another approach to developing benchmarks effect sizes in psychological research has been to

DISCUSS ESTIMATES OF THE ACTUAL AVERAGE EFFECT SIZE SEEN IN FIELDS OF PSYCHOLOGICAL RESEARCH

Systematic review

A Google scholar search of “average effect size” found 7 articles.

See C:\Users\fsingletonthorn\Documents\PhD\Dissertation Documents\List of papers providing effect size benchmarks.xlsx for tracking document

Cooper, H., & Findley, M. (1982). Expected Effect Sizes: Estimates for Statistical Power Analysis in Social Psychology

* Probably possible to just double the f statistics to get d for df = 1, maybe for greater ones as well

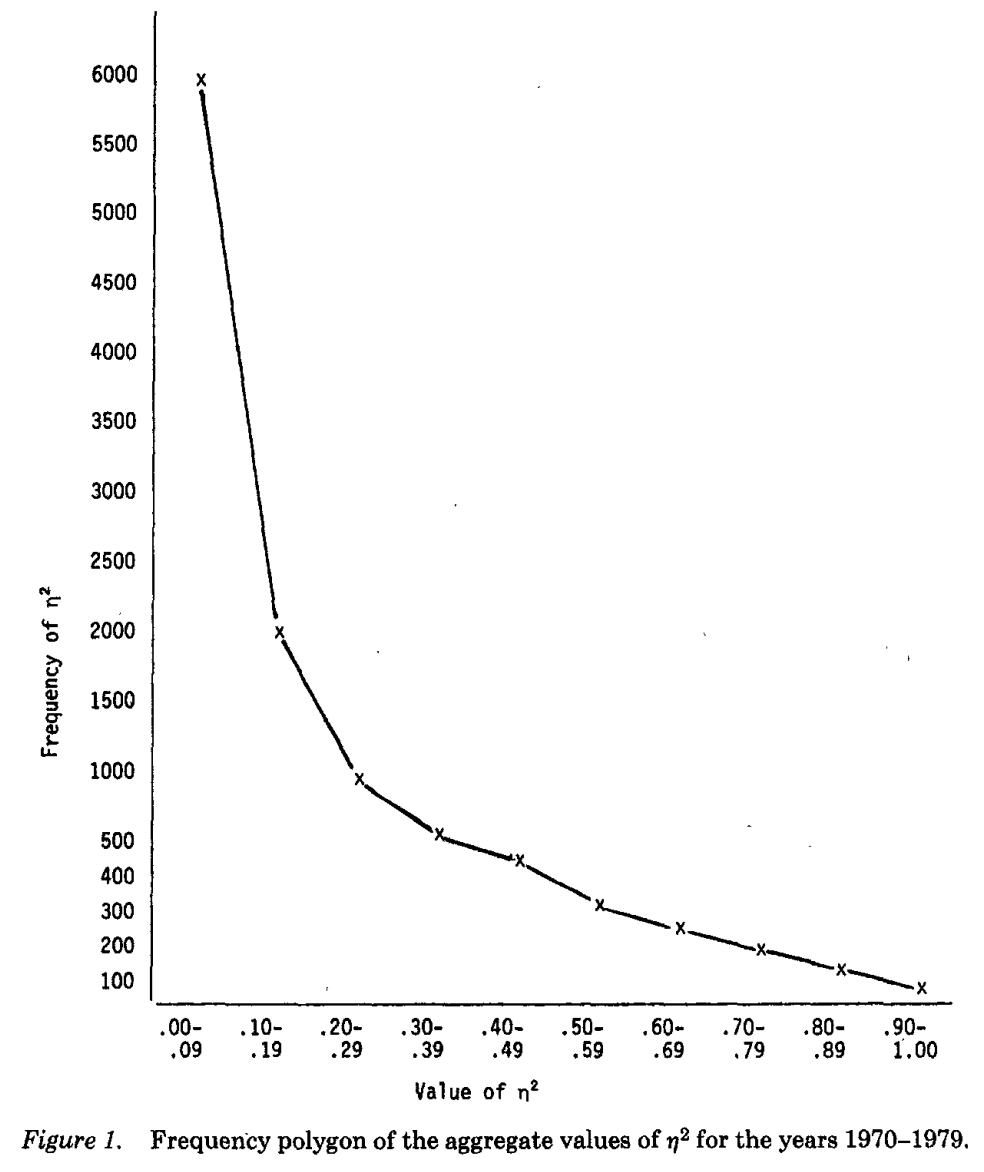


Figure [counselling] Frequency polygon of all univariate inferential statistics reported in the Journal of Counselling Psychology, 1970-1979 Reproduced from (Haase, Waechter, & Solomon, 1982)

4.1.4

[EffectSizeBenchmarksImage.R] FIX % location

Maybe explain what happens to the sampling distribution for the mean difference as the groups get further apart. Non-central ts, etc.

4.1.5

Attempts to estimate

Richard, Bond Jr., & Stokes-Zoota (2003) One hundred years of social psychology …

Is it reasonable to use this as a guide to determining sample sizes – no, not really, it’s going to be inflated. Attempts to correct for sample size… MAYBE DO THIS ???

MAYBE TRY TO ACCOUNT FOR PUBLICATION BIAS IN SOME WAY??

Other reasonable benchmarks? **Minimum clinically significant estimates?**

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

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

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.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159. doi:10.1037/0033-2909.112.1.155

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

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>

Haase, R. F., Waechter, D. M., & Solomon, G. S. (1982). How significant is a significant difference? Average effect size of research in counseling psychology. *Journal of Counseling Psychology, 29*(1), 58-65. doi:10.1037/0022-0167.29.1.58

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., & 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

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

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

Reiser, B., & Faraggi, D. (1999). Confidence Intervals for the Overlapping Coefficient: the Normal Equal Variance Case. *Journal of the Royal Statistical Society: Series D (The Statistician), 48*(3), 413-418. doi:10.1111/1467-9884.00199