**COPHAR**

**Pharmacological Perspective**

**2/1/19**

**MW review 2/10/19**

**DPL 10/8/19**

**Edit 090219**

**Null Hypothesis Significance Testing and Effect Sizes:**

**Can we 'effect' everything … or …anything?**

David P. Lovell1

1 St George’s Medical School, University of London,

Cranmer Terrace, London, SW17 0RE, UK

Version: 10th August 2019

**Running Title:** Significance Testing and Effect Sizes

**Keywords:** Reproducibility, Confidence Intervals, Significance, Effect size

**ABSTRACT**

The Null Hypothesis Significance Testing (NHST) paradigm is increasingly criticized. Estimation approaches such as point estimates and confidence intervals, while having limitations, provide better descriptions of results than P-values and statements about significance levels. Their use is supported by many statisticians. The *effect size approach* is an important part of power and sample size calculations at the experimental design stage and in meta-analysis and in the interpretation of the biological importance of study results. Care is needed, however, to ensure that such effect sizes are relevant for the endpoint. Effect sizes should not be used to interpret results without accompanying limits, such as confidence intervals. New methods, especially Bayesian approaches, are being developed; however, no single method provides a simple answer. Rather there is a need to improve researchers understanding of the complex issues underlying experimental design, statistical analysis and interpretation of results.

**Key Points**

* The null hypothesis significance test (NHST) approach is flawed and should be avoided where possible
* Estimation with measures of uncertainty e.g. point estimates and confidence intervals is a better and practical alternative
* Effect sizes are important for experimental design and meta-analysis
* The use of effect size measures to assess results in the absence of statistical analysis should be avoided
* Many new approaches including Bayesian methods are being proposed
* There is a need for researchers to be educated in the underlying concepts of experimental design and statistical analysis as well as learning the skills of carrying out statistical tests

**Introduction**

Nearly every pharmacologist has been exposed to aspects of the ongoing debate on the need for valid statistical analyses for testing the differences between treatments within an experiment to ensure reproducibility [1]. The journal *Nature* recently published a commentary entitled "Retire Statistical Significance" [2] together with a letter signed by over 800 statisticians which called for "…the entire concept of statistical significance to be abandoned". Specifically, this was "…for a stop to the use of *P* values in the conventional, dichotomous way — to decide whether a result refutes or supports a scientific hypothesis". This came at the same time as the American Statistical Association (ASA) produced a special issue with an introductory paper by Wasserstein et al [3] followed by 43 papers on what to do about statistics!

These are issues that are equally relevant for experiments carried out using cells, organs, animals or humans. Basically, any experiment that is designed to test a hypothesis should be well designed and powered and its variation minimized and/or controlled by the study design and statistical analysis [4].

To the practicing pharmacologist, there is a need for informed insight to help identify and understand the nuances of the statistical debate and to make the best decision in selecting the type of statistical approach for a series of experiments and how different analyses can best inform an experimental outcome. Among the issues under debate are:

the appropriateness of the traditional null hypothesis significance testing (NHST) paradigm,

whether identifying the size and biological importance of an effect is more important than whether the difference attains a threshold such as P < 0.05,

the use of effect sizes to help interpret and design studies and

whether the development of Bayesian statistical methods represents a realistic challenge to the more traditional frequentist approaches.

In this *Pharmacological Perspective*, the author attempts to provide some clarity on these topics.

**The Null Hypothesis**

It is now 25 years since Cohen introduced the NHST acronym in his paper "The Earth is Round P <0.05" [5]. It was intended to highlight how the NHST approach is severely affecting how scientific results are judged.

The main stumbling block is the concept of hypothesis testing. Superficially simple and deceptively useful for helping make difficult decisions - stop or proceed, worked or did not, good or bad - ideal for the decision maker yet fatally flawed. The concept of the Null Hypothesis, seductively similar to the Popper concept of the refutable hypothesis, turned out to be a clumsy hybrid of two strands of thought from the 1920s and 1930s (with a dash of Bayesian thinking thrown in) and the source of one of science's bitterest feuds. Fisher (developer amongst other things of null hypotheses and P-values [6]) and the duo of Pearson and Neyman (developers of confidence intervals, alternative hypotheses, Type 1 and Type 2 errors [7]) agreed on very little but would have agreed, once they realized how it was being used, that the NHST approach is a monstrous chimaera.

The Neyman-Pearson (NP) approach [7] is basically a straight 'accept or reject' with a fixed P-value as cut-off: above 0.05, accept and below 0.05, reject (irrespective of how low). Fisher only had a null hypothesis and the *P*-value was an indication of how strong the evidence was. (Fisher was very good with the quotable phrases but could also contradict himself). The 2 x 2 table of the possible results using the NP approach can be found in many introductory statistics text books or courses and forms the core of many basic statistical packages with their list of possible tests for carrying out such tests (often with 'cut and paste' phrases to include in reports).

Even before Cohen's paper there was trenchant criticism of the approach and this has gathered pace in subsequent years to the extent that it is difficult to find any professional statistician who will wholeheartedly support the NHST approach. Yet it persists, even flourishes in the scientific literature, together with its accomplices Dynamite Plunger / Coffee Pot graphics [8] and Amazon/TripAdvisor-like star ratings of significant findings in attempts to achieve in Tukey's words, 'statistical sanctification' [9]. Perhaps this could be manageable among the established journals but the thousands of new/fake/predatory journals appearing with no or limited editorial standards creates a major challenge [10].

The argument about *P*-values and significance levels took on a new urgency following the decision of the journal *Basic and Applied Social Psychology* (BASP) to ban P-values [11]. They were not the first: in the 1980s Rothman asked authors to remove references to statistical hypothesis testing and statistical significance from papers they submitted to the journal *Epidemiology* [12]. Partly spurred on by the BASP decision the ASA held a two-day meeting in 2016 and then published a statement on *P* values [13]. This statement put forward 6 principles and was accompanied by 23 supplementary commentaries which showed there was no real consensus although it concluded that there was "… no doubt that the 'best' estimate and CI approach is better by being more informative than the pure *P*-value". They pointed out that it "… can be equated to the NHST if one chooses to use it, although some use this equivalence as evidence of its inappropriateness".

In 2019, the ASA revisited the topic with a special issue in the *American Statistician* with an introductory paper by Wasserstein et al [3] who stated that "Significance tests and dichotomized *p*-values have turned many researchers into scientific snowbirds, trying to avoid dealing with uncertainty by escaping to a “happy place” where results are either statistically significant or not." They put forward 5 “don'ts". They stated, "In sum, “statistically significant”—don’t say it and don’t use it" and remember “ATOM” - “**A**ccept uncertainty. Be **t**houghtful, **o**pen, and **m**odest.”

Again, there was not complete consensus although there was probably an agreement that there were problems with the NHST approach even if not how to deal with it. Ioannidis argued against banning hypothesis testing because there are areas such as screening for genetic variants in Genome Wide Association Studies (GWAS) and quality control methods where binary decisions are useful. He and others argued that the NHST is so ingrained in the scientific method that it is not easily replaced and were concerned that replacing the approach could lead to the abandonment of statistical analysis in favor of subjective assessment. He argued that removing the 'Gatekeeper' of the *P*=0.05 threshold carries the risk that "…any result may be directly claimed to reflect an important signal or fit a pre-existing narrative" and he would rather see more stringent thresholds of 'significance'. Rules set beforehand are useful by limiting the potential for *post hoc* and subjective statistical inferences. Without such rules bias can be introduced but he stresses that in the case of bias “the only direct protection must come from standards for reproducible research” [14]. Others, like Cumming with his concept of *The New Statistics* [15] are critical of the problems of the P-value and are strong proponent of estimation and confidence intervals. They prefer focusing on effect sizes to dichotomous conclusions. Others argued that alternative approached such as Bayesian methods were to be preferred. Others that, while the NHST approach was adequate, the threshold for designating 'statistical significance' should be lowered, to say, *P* < 0.005 [16]

 So, after all these ‘don’ts, what to do? The ASA ‘collection’ came up with a range of replacements for the *P*-value: s-values, SGPVs, AnCred, mFPRs, BFBs, two-stage approaches using *P*-values and effect sizes, as well as various modifications such as procedures combining frequentist and Bayesian approaches, the use of statistical decision theory and amongst others, including the addition of a 'confidence index' to *P*-values. The developers of this [17] admit, however, that "Although simple on paper, requiring a confidence index would entail a profound overhaul of scientific and statistical practice." *Nature* [18] in an editorial called for education on statistical misconceptions and the need to consider aspects of 'uncertainty' other than just *P*-values and explore multiple ways of analyzing the data.

A practical implication of the NHST is that it does not give an estimate of the size of the effect of interest or the precision associated with it. Hence the increasing emphasis on estimates and confidence intervals. The case for moving from hypothesis testing - is there an effect? - to estimation using confidence intervals to show how big an effect is and how much uncertainty there is around the estimate has been made by many authors [19; 20]. Estimates and confidence intervals as opposed to hypothesis testing have for over 30 years been the method preferred by the International Committee of Medical Journal Editors for reporting results [21]. (The change occurred in the 1988 Third Edition <http://www.icmje.org/recommendations/archives/summary78-04.pdf>).

The definition of a confidence interval (CI) is, however, arcane [22]. Some critics point out that it still has an NHST concept underpinning it while Bayesian critics point to a logical fallacy and propose an alternative, the *credibility interval*. Unlike a CI, the credibility interval is dependent on the prior distribution that is specific to the situation. With CIs, the parameter is treated as a fixed value with the boundaries being random variables. For credibility intervals, the estimated parameter is treated as a random variable while the boundaries are considered fixed. Some proponents accept the NHST criticism but take the more pragmatic approach that the estimate and confidence interval provide much more practical information than the *P*-value. The development of 'estimation graphics' [23] which provide a resource for presenting a visualization of results based upon estimation provides a further advantage. Put simply - *estimates and confidence intervals are better than the NHST.*This approach has many advocates [19]. A key point is that it is not the statistical significance which is important but that the effect size is sufficiently large to be interesting and lead to further work.

**Effect sizes**

Cumming [15] described an effect as 'anything we might be interested in' and an effect size 'the size of anything that may be of interest'. There are, though, numerous other definitions with different meanings as the concept of the effect size has evolved [24]. Kelley & Preacher [25] noted the confusion in the literature on what an effect size is and proposed a definition for 'effect size' as "…a quantitative reflection of the magnitude of some phenomenon that is used for the purpose of addressing a question of interest" and discussed how it can be reported and interpreted.

In practice, the term can be used in two ways: either as observed effect sizes – such as the results of experiments - or as planned effect sizes. Observed effect sizes are the quantification of differences, for instance between the means of, for example, a control and a treated group. The CI associated with the effect size should usually be included in the reporting. Planned effect sizes are a key part of the design of experiments particularly for power and sample size estimation. Observed effect sizes may be referred to as 'broad' while planned effects as 'narrow' although this usage is not uniform (<https://transparentstatistics.org/2018/07/05/meanings-effect-size/>. The recent introduction of terms such as *estimands* [26], the target of estimation (as distinct from the method of estimation – the *estimator*) is, in part, an attempt to clarify and pre-define what is measured in a study and to remove the potential for subjective post-hoc interpretation of results

At the analysis stage an observed effect size is a point estimate of, for instance, the difference of two means with an appropriate CI. This provides a 'best' estimate of the size of the result and a measure of the uncertainty of the estimate. Planned effect sizes are theoretical values defined at the planning stage and should represent differences which are considered biologically or clinically important between, for example, groups and that it is hoped to be able to detect if they exist. Effect sizes are a key aspect of both the planning of experiments, especially for clinical trials, and for the combination of studies in meta-analyses. There are two broad ways of reporting them: either as absolute values based upon the original value or as relative values based upon standardized values such as expressed relative to the standard deviation. The latter provides a uniform measure suitable for inclusion in meta-analyses.

In some fields an effect size approach has been proposed as an alternative to the NHST [27]. Nakagawa & Cuthill [28] termed the effect size approach "effective thinking" which refers to "... the philosophy of placing emphasis on the interpretation of overall effect size in terms of biological importance rather than statistical significance". An 'effect size movement' has developed in the field of educational and psychology research [25, 29] leading to many journals in these fields making the reporting of effect sizes mandatory [24].The approach can in some cases be interpreted as replacing statistical analysis completely by an expert, but subjective. assessment of the effect sizes or by relating the results to pre-defined effect sizes and referring to substantive as opposed to statistical significance [25]. This approach is unsatisfactory and is one fear of Ioannides [14] and others that criticism of NHST approach can lead to even less satisfactory approaches being used. Kelley & Preacher [25] argued strongly that interval estimates especially CIs, should accompany effect sizes, with a failure to do so being a major weakness. Effect sizes may be based upon effects found in previous studies or they may be 'global' effects like the Cohen criteria, but in both instances they may be problematic.

A criticism of the widespread use of effect size criterion such as 'medium' using Cohen's *d* (see below) is that they can be unsuitable for assessments across studies and/or endpoints where they may not be relevant. The effect sizes need to be relevant for the particular research question. Specific endpoints need adequate data to be able define, for example, small, medium or large effects for the biological contexts of a specific discipline. This may not always be available. A further complication is that the standard deviation may not be an appropriate statistic to use for standardization when the data are not normally distributed.

**Power and sample size calculations**

Despite the limitations of the NHST one aspect of it is widely used (and requested by grant panels and ethics review boards/committees). This is the concept of power and sample size calculations at the planning stages of experimental studies. Much of the development of the power and sample size methodology was carried out by Cohen [30] in the late 20th Century although it built on earlier work [24). He was a strong proponent of effect sizes over significance testing: "The primary product of a research inquiry is one or more measures of effect size, not P values."

Although many statisticians have reservations about the NHST they are relatively comfortable in using concepts from it to carry out such power and sample size calculations. These are based upon estimates of the sizes of effects expected to be detected, the variability of the study material together with the defined α (Type 1) and β (Type 2) error rates. Power (or (1- β) (a term first used by Neyman & Pearson in 1933[7]) is the probability of detecting an effect, given that the effect is really there. Alternatively, it is the probability of rejecting the null hypothesis when this is false and thus not committing a Type II error.

The calculations for simple designs (and, in the case of some packages, more complex designs) can be obtained using a number of different specialised software packages (including PASS, nQuery Advisor, GPower). A range of websites can also carry out simple calculations. (There is, therefore, no longer a need to get a calculator out to use standard formulae!) Some of the standard statistical packages have options for sample size calculation. Packages in R, a language for statistical computing [31], can also be used. Several books have tables for deriving sample sizes [32], and Apps have been developed for smartphones [33].

The value of power used in the analyses is often 0.8 (80%). In clinical trials, increasingly, 90%. There does not seem to be a specific recommendation for the choice of these percentages. However, Sullivan & Feinn [34] noted that a β error of 0.2 (equivalent to a power of 80%) was chosen by Cohen because he argued that the α error was (4 times) more serious than the β error, hence the β error of 0.20.

The inputs to the calculation of the sample sizes required for a two-group experiment for a quantitative measure are the alpha (α) level, the power (1- β), whether the statistical test is a one- or two-sided, the effect size and some measure of the variability of the measure, usually the SD. Alternatively, the sample size can be included to provide an estimate of the power of a specific design. Similar inputs for binary data are required with the control and treated proportions replacing the effect size and the SD.

Identifying an appropriate effect size can be a challenging task. A key aspect is that this should be an effect of a size that would be biologically important/relevant such that if it was found it would influence, for instance, how further studies would be carried out. It is pre-defined at the planning stage of the project. This variable/concept goes under several names, such as "the difference worth detecting" [35]. One from the clinical sector is the "Minimal clinically important difference" (MCID [36]).The MCID is referred to in different ways: clinically relevant difference*,* minimum important difference, minimum clinically important difference (most commonly used), cut-off value of success, minimum clinical important change, MDC, meaningful differences, and smallest detectable difference [37].

**Effect sizes in power calculations**

In the 1980s, Cohen [29] investigated the sample sizes needed for various types of simple designs and published tables associated with them. He suggested an effect size termed *d*, which was standardized by being based upon a size in SD units. In a two-group study this was the difference in group means divided by the within group SD. In the case of quantitative comparisons between two groups he chose three measures of *d*: 0.2, 0.5 and 0.8 SD units and termed them: small, medium and large differences. The values were chosen based on the degree of overlap there would be between two normal distributions and based on results from research in the social sciences, He used the distributions of schoolchildren heights at different ages to illustrate the different effect sizes. Rosnow & Rosenthal [38] added a 1.3 SD effect size which they termed 'very large' while Sawilowsky, [39] added 0.01 as 'very small' and 1.2 and 2.0 as 'very large' and 'huge', respectively.

Cohen's approach allows a simple 'rule of thumb'/informal calculation of sample sizes for comparisons between two groups. Sample sizes of approximately 16 for each group (call this 20 for a 'round number' sample size) have 80% power to detect a 1 standard deviation (1 SD) difference between the two means at the two-sided  = 0.05 level. The approximation is based on a formula for calculating sample sizes for two group studies which contains the term, n= 16s2/m2 where s is the standard deviation and m is the difference between the two means in standardized values. The term in the equation denotes that there is an approximately 4-fold increase in the sample size needed with each halving of the effect size. So approximate sample sizes for increasing effect sizes expressed in standard deviation (SD) units: 0.125 SD, 0.25 SD, 0.5 SD, 1.0 SD and 2.0 SD units are, 1024, 256, 64, 16 and 4 respectively. The comparable sample sizes obtained using sample size software gives n values of 1006, 253, 64, 17 and 6 respectively. Another simple 'rule of thumb' is that the 'N=3 rule' prevalent in some areas of biology has about 80% power to detect a 3 SD effect (two-sided  =0.05).

Cohen subsequently, in a short paper ,'Power Primer' [40], set out operationally defined effect size (ES) indices for various other types of statistical tests including criteria for correlation coefficients, anovas and tests of proportions and provided sample sizes associated with the ESs for these tests. It should be noted, though, that Cohen's 'conventions' have different meanings for different tests and that his criteria were derived from the social science literature and may not always be relevant to other fields.

The standardized effect size approach, although useful, has its critics. Glass *et al* [41] were particularly critical of the approach believing it to be dissociated from a context of decision and comparative value. Lenth [42] called the definition of small, medium and large ESs the 'T-shirt approach' and argued that these “canned” effect sizes provided an uncritical consideration of whether the effect size was biologically important. He was concerned that they risked being used to avoid difficult issues such as deciding on the absolute effect size and variability to include in the calculation so, in effect, resulting in a way of defining large, medium, or small sample sizes without relating these to the absolute effect. Others extended the T-shirt 'analogy' by calling ESs of 1 SD as XL, 1.5 SD XXL and 2 SD as XXXL while Ellis [43] provided another way of defining the size of a standardized ES with his cartoon "Result Whacker" that require a standardized mean difference (d) or strength of association (r) to be input <https://www.polyu.edu.hk/mm/effectsizefaqs/calculator/result.html>.

**Effect sizes in more complex designs**

A complication with determining power for more complicated experimental designs is the use of null and alternative hypotheses in the calculations. Moving from the simple 'two group' design to multiple groups introduces more complexity in the definition of the alternative hypothesis. For instance, in the case of a four-group design, is the alternative hypothesis that the four groups differ or that there is a clear dose-response relationship? The power to address these different hypotheses with the same sample sizes will be different. The software in many statistical packages can handle the simple designs and may include options for more complex designs. However, the user needs to be aware of what these methods may require as input. For instance, in a one-way anova with, say, four groups, the package Minitab [45] will give sample sizes for the 'omnibus' test of the Null Hypotheses that all the groups are from the same population based upon the range between the highest and lowest mean using the pooled within group standard deviation (SD) while G-Power [45] will require the means for each group to be used in the estimate of the ES. The two methods will give different estimates for the sample sizes. The assumptions are more complex still for more advanced designs such as the widely used factorial designs [46]. It can also be difficult to identify exactly what the ESs relate to when the hypotheses become more complicated with the more complex designs.

Although approximate formulae are used in some programs, the methodology for deriving the Power/Sample Size estimates for more complex designs involves using matrix algebra on the non-central distributions associated with the calculations. This is a complex area and few 'closed form equations' exist [42]. A closed form equation is one where the solution can be obtained just by 'plugging' values into a formula as opposed to an open form equation where iteration is necessary.

One approach to these types of more complex designs is the use of simulations and scenarios. This approach is now widely used in the pharmaceutical industry for the simulation of clinical trials and to pre-test a proposed Statistical Analysis Plan (SAP) and to provide templates for the reporting of results such that the reporting stage becomes a case of filling in the blanks in the pre-specified tables. One complication of this approach is that is can be quite time-consuming to identify all the possible options, simulate them and assimilate the findings.

**Estimates of variability**

An important complication is that the ES sizes are affected by the size of the estimate of the variability in the calculation. Estimates of variability are usually derived from historical data or, in some cases, pilot studies. However, estimates of variability of statistics such as variances and SD are very imprecise and have wide confidence intervals (CIs). For example, if a measure has a within-group standard deviation of 100 units then the 95% CI for a SD of this size from a sample of 10 is from 69 to 183 units while for a sample of 100 the 95% CI would be from 88 to 116 units (notice that, particularly for small sample sizes, the CI is not symmetrical). This illustrates the need for good estimates of the SD by showing that if the 'true' SD is actually twice as big as the 'estimated' SD then the standardized effect size (*d*) will, in practice, be halved and the estimate of the group sizes would need to be 4 times larger. Estimates of SD based upon n =10 could be approximately 80% higher or 30% lower. Even with sample sizes of 100 the 95% CI would be 16% higher or 12% lower.

This indicates that pilot studies based upon small sizes will not provide precise enough estimates of the sample sizes needed to achieve the planned power. Estimates of SDs based upon relatively small samples, e.g. less than 30, have wide confidence intervals and will result in imprecision in the estimated sample size. The use of standardized effect sizes does not get around the problem because no estimate of the precision of the SD for the measure is included in the derivation of the effect size. Whitehead *et al* [47] describe approaches to estimating sample sizes for pilot studies based upon: i) estimates of the SD based upon either the 85% or 95% upper confidence limit (UCL) of the estimate of variance) or ii) a non-central t-distribution (a method which takes into account that the variance estimate is based upon a sample).

**Retrospective Power Analysis**

A retrospective power calculation is where a power calculation is carried out by including the observed results of the study in a power/sample size calculation. The overall advice is "Don't do them!" They are, in effect, just another way of carrying out a statistical test and reporting that it would have been statistically significant if only the sample sizes had been bigger. Calculating the retrospective power (by doing the power calculation using the actual study results) is just another way of formulating the *P*-value. It is commonly seen, but is very wrong. Lenth [48] has been strongly opposed to the use of retrospective power analyses. It should, however, not be confused with *post hoc* power calculations where the objective is to assess whether the published data had a reasonable chance of rejecting an incorrect null hypothesis [45].

**Relating power and effect size to statistical significance**

The result of an experiment will produce an observed effect size which is a single observation from an underlying distribution of a 'theoretical' effect sizes. It is not possible to know whether this single sample represents the 'true' value or is a low value from a distribution of a larger effect size or, conversely, a high value from a distribution of a smaller effect size. However, a consequence is that some results can be identified as statistically significant because their *P*-values are below the P=0.05 Neyman-Pearson threshold but which will have observed effect sizes below the effect size included in a power/sample size calculation. In such cases, while a statistically significant result has occurred, the size of the effect is not biologically important.

Figure 1 illustrates the difference between the *P*-value and the confidence interval/ effect size approaches. The black symbols in the figure show the size of differences between the means of two groups that would be just significant (*P*=0.05) in a two-sided test when groups are increased from 5 per group to 100. The height of the vertical lines above zero joining the black circles shows the difference (in SD units) between the two groups which would be just significant at *P*=0.05 for different group sizes. The height from zero is also equivalent to the 95% CI of the difference between the two means. This illustrates the close relationship between the CI and the significance test- a point of criticism by those who prefer alternative approaches. The effect size associated with a power of 50% is effectively equivalent to the 'just significant difference' at the *P*-value of 0.05. The observed effect size which produces a result of *P*=0.05 would, if entered into a power calculation, have close to 50% power. The three horizontal dashed line represent the Cohen effect sizes, *d*, in SD units of, from top to bottom: 0.8, 'large'; 0.5, 'medium' and 0.2 'small'.

An important point is while the significance level (P=0.05) remains constant, the observed effect sizes differ greatly between studies with different sample sizes [49]. The 95% CI of 1.46 SD units for sample sizes of 5 is shown is shown in the figure together with, on the right-hand side, the 95%CI of 0.28 SD units for sample sizes of 100. This is over five times smaller but with the same *P* value. It should be is noted that the effect sizes when the group sizes are in the region of 1000 are in the range of 0.125 SDs.

The second line, with red symbols, shows the effect sizes associated with studies designed with 80% power. This give an indication of the effect size implicitly (and, hopefully explicitly) considered biologically important at the time a study is designed. In contrast, the effect size (black symbols) which is just detected as significant at *P*=0.05 is effectively that associated with 50% power. The difference between the two lines indicates where effects, although being deemed be 'statistically significant', are below the predefined effect size and, therefore, should not be considered biologically important. The gap between the lines shows that effects up to 30% lower than those defined as 'biologically important' sizes can be detected as statistically significant using the *P*=0.05 threshold to declare a significant event. The 'gap' appears particularly pronounced in Figure 1 when studies with very low small sample sizes are considered.

**Limitations of effect size and power calculations**

Power calculations using effect sizes are useful but are not a panacea. They are also based upon the same assumptions that underlie the NHST which is that the experimental units will be randomly assigned and be independent of one another. If not, there is the potential for biases and false positive results to occur especially with observational studies and where pseudo-replication is introduced. Lenth [42] has argued that carrying out Power/Sample Size calculations requires that science is put before statistics by which he meant that there needs to be a "serious discussion of study goals and effects of clinical importance, on the actual scale of measurement".

There is always the need for a pragmatic approach to power/sample size calculations. These should form an integral part of the design process for experiments. The Power and Sample Size calculations are easy (too easy?) to carry out (there is much software available) if the design is simple such as with comparisons between two groups. Such calculations can be helpful in assessing how realistic the study is, given the possible economic and ethical constraints. It is important, though, that they should not be too prescriptive.

The implications of the inherent variability of estimates of variability means that over-precise estimates of sample sizes should be avoided: arguments over whether 'n' should be 4 or 5 are irrelevant. Certainly, sample sizes to one (or more) decimal places should be avoided. Far more important is that the power calculation provides evidence that some thought/ planning has gone into the design of studies. One must also be aware of the shadow of the NHST over the methodology. Remember that the objective should be to estimate the size of the observed effects together with the limits associated with them rather than whether the results fall one side or other of an arbitrary threshold.

**Use of effect sizes in meta-analysis**

Although the researcher, understandably, often has more interest in their own studies than the wider field, the scientific community, as a whole, is becoming increasing aware of the need for a 'synthesis' of research through techniques such as systematic review and meta-analysis [50, 51]. Meta-analysisanalysis is "an effect-size based review of research that combines results from different studies on the same topic in order to draw general conclusions by estimating the central tendency and variability in effect sizes across these studies" [28]. A major advantage of an effect size approach is that standardized effect sizes are degrees of effects that can be compared across studies. Standardized effects sizes, with confidence intervals, derived from the peer-reviewed literature and which have not been assessed through the lens/sieve of multiple comparison methods are crucial. Nakagawa & Cuthill [28] state "In meta-analysis, presentation of effect statistics and their CIs is mandatory" and link it closely to Ioannides' view [52] that "Eventually, all research (both primary and meta-analytic) can be seen as a large, ongoing, cumulative meta-analysis”.

Although standardized effect sizes are widely used for meta-analyses and in their interpretation, there is an expectation that researchers will also provide sufficient information from the unstandardized summary of the results to allow standardized results to be calculated when a meta-analysis is carried out.

**Use in model building**

Another area where effect size methods are used is in the field of statistical model building. Areas of research investigating observational studies which fit forms of regression models to observational data include, for instance, epidemiology or education research in the social sciences. Sample sizes can be determined using software such as GPower [45] to estimate the sample sizes need to test whether including extra or fewer parameters in a model will improve the fit of the model as measured by statistics such as R2 values. This approach links to the more sophisticated experimental designs and analyses based around the generalization of the analysis of variance methodology into the General Linear Model procedures and onwards into other more advanced modelling approaches.

**Bayesian approaches**

Users of these methods should also remember that there is an alternative statistical philosophy, Bayesian, which can both challenge and reinforce the conclusion drawn from the frequentist strand of statistical science. Frequentists and Bayesians have fundamentally different philosophies and methods. These differences provoked often bitter debates between the two schools during the 20th Century. Although differences remain, the 'taking of sides' is now much less pronounced, perhaps even a rapprochement, with many statisticians appreciating the strengths and weaknesses of both approaches. Frequentists base their analyses on introducing hypothetical populations and then trying to estimate parameters using data from samples while the Bayesians introduce extra information, such as results of previous studies or expert but subjective estimates to create a 'degree of belief' or 'prior probability' into the analysis. The choice of prior probabilities for inclusion in Bayesian analyses is one area of contention in the use of the methodology.

Although, dating back to its eponymous clergyman developer in the 18th Century, Bayesian approaches had a greater academic than practical impact until the latter part of the 20th century. At that point increasing computer power made it possible to develop specialist algorithms and software so that the potential of Bayesian approaches could be applied to more routine applications. Software such as BUGS (Bayesian inference Using Gibbs Sampler) and now available as WinBugs and OpenBugs, [53], numerous packages within the R the computing language and software environment for statistical computing have opened up opportunities while the more established packages such as SAS and SPSS are now including Bayesian analyses in their newer versions.

The increasing acceptance of Bayesian approaches has led to alternative approaches to statistical testing. In response to the crisis in reproducibility there has been considerable emphasis on the use, and mis-use of p-values, and significance testing resulting in the ASA statement on p-values [13] This has led, in turn to a proposal based upon a Bayesian approach to use a more stringent threshold *P*-value of 0.005 to denote statistically significant findings [16]. This approach has met considerable criticism [54].

**Concluding remarks**

The NHST approach has long been recognized as having many limitations and has recently come under renewed criticism from many statisticians. It has some uses where clear binary decisions have to be based solely on the data. There is, also, an argument that it is now so widely used that attempts to replace it risks making the problem worse because decisions may be made solely using subjective judgement. Estimation approaches such as point estimates and confidence intervals, while having some limitations, provide a pragmatic alternative approach by providing useful information on the size of effects and the uncertainty associated with them than P-values and statements about statistical significance.

Aspects of the estimation process are called by some the effect size approach. Effect sizes are useful in the design of studies and in meta-analysis. The use of 'canned' effect sizes, such as Cohen's, in the assessment of results can be criticized because the biological underpinning relating these effect sizes to the importance of the effect may not be there. The use of effect size to assess results in the absence of limits, such as confidence intervals, is misguided and should not be encouraged.

A large number of alternative methods are being suggested as can be seen in the responses to the recent ASA statements. Bayesian methods, for instance, are likely to increase in use. All have their proponents, and all, probably, also their critics. Some contributors to the ASA special issue argued that study designs and statistical analysis plans should be pre-specified, even publicly pre-registered, before the start of a study. Plans should be strictly followed during the conduct of the study with any changes tracked over time. Exploratory studies should be explicitly distinguishing from confirmatory studies and statistical inferences restricted "…to confirmatory analyses for which the study design and statistical analysis plan are pre-specified prior to, and strictly adhered to during, data acquisition" [55].

However, no single method is likely to be able to provide a solution which replaces the NHST. Instead there needs to be a major and continuing effort to re-educate researchers, authors, reviewers and editors to reverse the continuing use of the inappropriate NHST paradigm. They need to appreciate that the objective of a statistical analysis should not be to provide certainty to any decision but rather to provide some bounds on the degree of the uncertainty in results. They need to appreciate the role of the statistical methods is to provide estimates of the size of effects. Rather than present the result as a binary/dichotomous, significant versus non-significant, positive / negative result conclusion they should instead provide some measure of the degree of the uncertainty in the results which can aid, together with other relevant information in the interpretation of a result. "Statistics is both the science of uncertainty and the technology of extracting information from data" [56]. It is important that future and current scientists learn more about the subtleties of experimental design, statistical methods and the interpretation of results as well as the 'core skills' to put data though formulaic tests using statistical packages

**References**

1. [Goodman, S.N](https://www.ncbi.nlm.nih.gov/pubmed/?term=Goodman%20SN%5BAuthor%5D&cauthor=true&cauthor_uid=27252173)., [Fanelli, D](https://www.ncbi.nlm.nih.gov/pubmed/?term=Fanelli%20D%5BAuthor%5D&cauthor=true&cauthor_uid=27252173), & [Ioannidis. J.P](https://www.ncbi.nlm.nih.gov/pubmed/?term=Ioannidis%20JP%5BAuthor%5D&cauthor=true&cauthor_uid=27252173). (2016) What does research reproducibility mean? Science Translational Medicine 8 1-6
2. Amrhein, V., Greenland, S. & McShane, B. (2019) Retire statistical significance. Nature 57 305-307
3. Wasserstein, R. L, Shchirm, A.L. & Lazar, N. A. (2019) Moving to a World Beyond “p < 0.05" American. Statistician 73:sup1 1-19.
4. Ioannidis, J.P., Greenland, S., Hlatky, M.A., Khoury, M.J., Macleod, M.R., Moher, D., Schulz, K.F. & Tibshirani, R. (2014) Research: increasing value, reducing waste 2) Increasing value and reducing waste in research design, conduct, and analysis. Lancet 383 166–175
5. Cohen, J. (1994) The Earth Is Round (p < .05) American Psychologist. 49 997-1003
6. Salsburg, D. (2001) The lady tasting tea. W, H, Freeman and Company, New York.
7. Neyman, J. & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. Philos. Trans. R. Soc. Lond. A 231 289-337
8. Drummond, G.B. & Vowler, S.L. (2011) Show the data, don’t conceal them British Journal of Pharmacology 163 208–210
9. Tukey, J. W. (1969.) Analyzing data: Sanctification or detective work? American Psychologist, 24 83 - 91
10. Lemming J (2017) Science's fake journal epidemic. Predatory publishers, peerless reviews and those who fight against the destruction of the scientific approach. Available from: http://blogs.nature.com/naturejobs/2017/06/26/sciences-fake-journal-epidemic/.
11. Trafimow, D. (2014) Editorial . Basic and Applied Social Psychology 36 1 –2
12. Fidler, F., Thomason, N., Cumming, G., Finch, S. & Leeman, J. (2004) Editors can lead researchers to confidence intervals, but can’t make them think: Statistical reform lessons from medicine. Psychol Sci.15 119-126
13. Wasserstein R. L. &. & Lazar, N (2016) The ASA's Statement on p-Values: Context, Process, and Purpose. American. Statistician 70 129–133
14. Ioannidis, J.P.A. (2019) The importance of predefined rules and prespecified statistical analyses do not abandon significance. JAMA 321 2067-2067
15. Cumming, G. (2012). Understanding The New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis. New York: Routledge
16. Benjamin, D.J. Berger J.O., Johannesson M, Nosek B.A., Wagenmakers, E-J, Berk, R., Bollen, K.A., Brembs, B., Brown, K., Camerer, C., s et al (2018) Redefine statistical significance. Nature Human Behaviour 2 6–10.
17. Goodman, S.H. (2019) Why is Getting Rid of P-Values So Hard? Musings on Science and Statistics. American Statistician 73 26-30
18. Nature (2019) Editorial: Significant debate. Nature 567 283
19. Gardner, M.A. & Altman, D.G. (1986) Confidence intervals rather than P values: estimation rather than hypothesis testing. Br Med J. 292 746–750
20. Altman, D.G. (2000) Statistics with Confidence: Confidence Intervals and Statistical Guidelines, 2nd Edition London: Wiley-Blackwell
21. International Committee of Medical Journal Editors (2018) Recommendations for the Conduct, Reporting, Editing and Publication of Scholarly Work in Medical Journals. <http://www.icmje.org/icmje-recommendations.pdf>
22. Morey, R.D., Hoekstra, R., Rouder, J.N., Lee, M.D. & Wagenmakers, E-J. (2016) The fallacy of placing confidence in confidence intervals. Psychon Bull Rev 23 103–123
23. Ho, J., Tumkaya, T., Aryal, S., Choi, H. & Claridge-Chang. A. (2019) Moving beyond P values: data analysis with estimation graphics. Nature Methods 16 565–566
24. Huberty, C.J. (2002) A History of Effect Size Indices. Educational and Psychological Measurement. 62 227-240
25. Kelly, K. & Preacher, J.K. (2012) On effect sizes. Psychological Methods 17 137–152
26. Akacha, M.F., Ohlssen, East Hanover, NJ, D., Rosenkranz, G. a & Schmidli, H (2017) Estimands and Their Role in Clinical Trials. Statistics in Biopharmaceutical Research 9 268-271
27. Ferguson, C.J. (2009) An Effect Size Primer: A Guide for Clinicians and Researchers. Professional Psychology: Research and Practice 30 532-538.
28. Nakagawa, S. & Cuthill, I.C. (2007) Effect size, confidence interval and statistical significance: a practical guide for biologists. Biological Rev Cambridge Philosophical Society. 82 591–605
29. Robinson, D. H., Whittaker, T. A., Williams, N. J., & Beretvas, S. N. (2003). It’s not effect sizes so much as comments about their magnitude that mislead readers. J Experimental Education. 72 51–64
30. Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd Ed.). Hillsdale, NJ: Lawrence Earlbaum Associates
31. R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL [http://www.R-project.org.](http://www.r-project.org./)
32. Machin, D., Campbell, M.J., Tan, S-B. & Tan, S-H. (2008) Sample Size Tables for Clinical Studies, 3rd Edition BMJ Books
33. Flight, L.G. & Julious, S.A. (2015) Practical guide to sample size calculations: an Introduction. Pharmaceutical Statistics 15 68–74
34. Sullivan, G.M. & Feinn, R. (2012) Using Effect Size—or Why the P Value Is Not Enough Journal of Graduate Medical Education 4 279-282.
35. Jones SR, Carley S, Harrison M. (2003) An introduction to power and sample size estimation. *Emerg Med J*. 20:453-458.
36. Jaeschke R, Singer J, Guyatt GH. (1989) Measurement of health status. Ascertaining the minimal clinically important difference. Control Clin Trials. 10:407-415
37. Coretti, S., Ruggeri, M. & McNamee P. (2014) The minimum clinically important difference for EQ-5D index: a critical review. Expert Rev Pharmacoecon Outcomes Res.14 221-233
38. Rosnow, R. L., & Rosenthal, R. (1996). Computing contrasts, effect sizes, and counternulls on other people's published data: General procedures for research consumers. Psychological Methods 1 331-340.
39. Sawilowsky, S (2009). New effect size rules of thumb. J Modern Applied Statistical Methods. 8 467–474
40. Cohen, J. (1992). A power primer. Psychological Bulletin. 112 155–159
41. Glass, G. V., McGraw, B., & Smith, M. L. (1981). Meta-analysis in social research. Sage.
42. Lenth R.V. (2007) Statistical power calculations. J Anim Sci. 85 E24–E29
43. Ellis, P D. (2010). The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis, and the Interpretation of Research Results. Cambridge University Press
44. Minitab 17 Statistical Software (2010). State College, PA: Minitab, Inc. (www.minitab.com)
45. 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 39 175-191
46. Wu, C.F.J. & Hamada, M.S. (2009) Experiments: Planning, Analysis, and Optimization. (2nd Ed.), Wiley.
47. Whitehead, A. L., Julious, S. A., Cooper, C. L. and Campbell, M. J. (2015) Estimating the sample size for a pilot randomised trial to minimise the overall trial sample size for the external pilot and main trial for a continuous outcome variable. Stat Methods Med Res 25 1057-1073
48. Lenth R.V. (2001) Some practical guidelines for effective sample size determination. Am Stat. 55 187–193
49. Cohen, J (1990) Things I Have Learned (So Far) American Psychologist 45 1304-1312
50. Glass G. V (1976). Primary, secondary, and meta-analysis of research. Educational Researcher 5 3–8
51. Cochrane Collaboration (2011) Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]. Higgins JPT, Green S (editors).
52. Ioannidis, J.P. (2010) Meta-research: the art of getting it wrong. Res Synth Methods 3 169–184
53. Lunn, D., Spiegelhalter, D., Thomas, A. and Best, N. (2009) The BUGS project: Evolution, critique and future directions (with discussion), Statistics in Medicine **28** 3049--3082
54. Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E. & Zwaan, R. A. (2017). Justify Your Alpha. Nature Human Behav 2 168-171
55. Tong.C. (2019) Statistical Inference Enables Bad Science; Statistical Thinking Enables Good Science. The American Statistician 73 246-261
56. Hand, D.J. (2009) Modern statistics: the myth and the magic. J. R. Statist. Soc. A 172 287–306

**Figure 1. Plot of Effect size v. Group size**



A plot of the effect sizes associated with a *just significant difference* (JSD) (P=0.05) in a two-sample two-sided test (black symbols) with increasing group sizes. The line with red symbols shows the effect sizes associated with 80% power in a two-sample two-sided test with alpha =0.05. The three dashed lines represent Cohen's effect sizes, *d*, in SD units of, from top to bottom: 0.8, 'large'; 0.5, 'medium' and 0.2 'small'. The error bars represent the 95% confidence interval for the effect size for group sizes of 5 (left) and 100 (right).