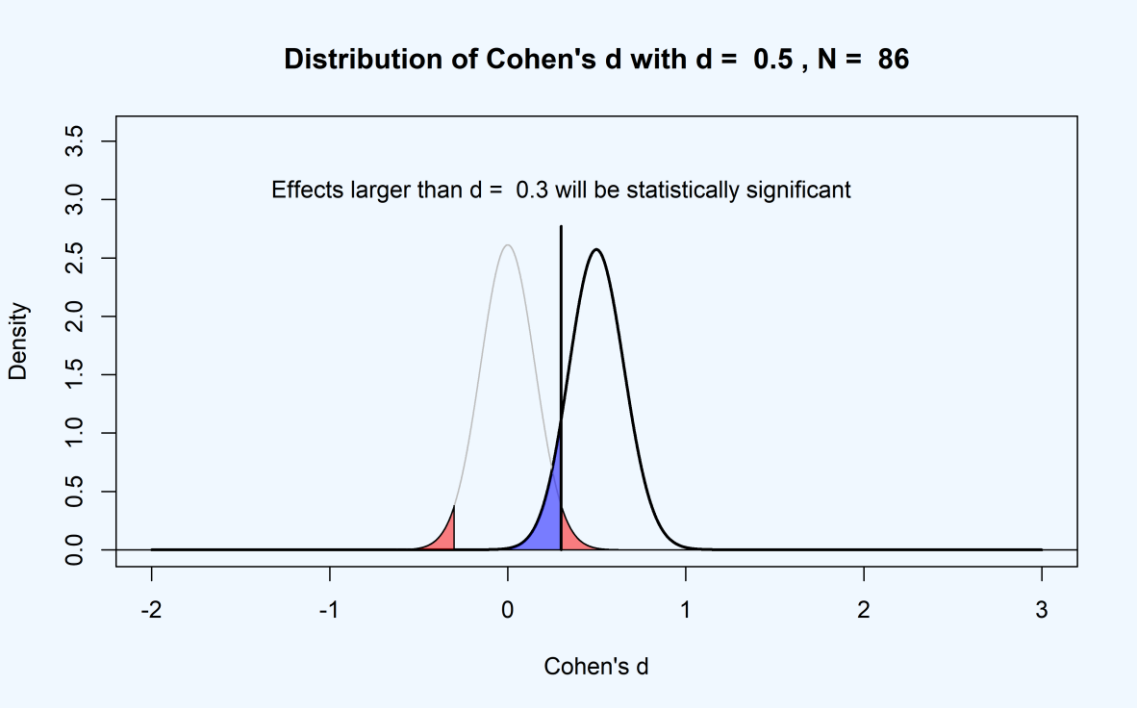
Set Your Smallest Effect Size of Interest Based on Resources

When designing a study, you need to justify the sample size you aim to collect. If one of your goals is to observe a *p*-value lower than the alpha level (e.g., 0.05), a commonly used justification for the sample size is a power analysis. A power analysis tells you the probability of observing a statistically significant effect, based on a specific sample size, alpha level, and true effect size.

A power analysis is performed based on some assumption about the effect size you want to be able to detect with reasonably high probability. When you would like to be able to observe an effect with a standardized mean difference in Cohen's d of 0.5 in an independent two-tailed *t*-test, and you use an alpha level of 0.05, you will have 90% power with 86 participants in each group. What this means, is that we can calculate the expected distribution of observed effect sizes when d = 0.5 and n = 86. In this case, only 10% of the distribution of effects sizes you can expect falls below the critical value required to get a $p < 0.05$ in an independent *t*-test.

In the figure below, the power analysis is visualized by plotting the distribution of Cohen's d given 86 participants per group when the true effect size is 0 (or the null-hypothesis is true), and when d = 0.5. The blue area contains 10% of the distribution when the true effect size is 0.5, and visualizes the 10% of results you can expect to observe in the long run that would yield $p > .05$.



You've probably seen such graphs before (indeed, G*power, a widely used power analysis software, provides these graphs as output). The only thing I have done is transform the *t*-value distribution that is commonly used in these graphs, and calculated the distribution of Cohen's d. This is a straightforward transformation (based on the *non-centrality parameter*). Given a test with 86 participants in each group, and an alpha level of 5%, only *t*-tests which yield a $t \geq 1.974$ will be statistically significant. In other words, $t = 1.974$ is the **critical *t*-value** (you might have learned that as the sample size becomes large enough, the critical *t*-value goes to 1.96).

Because I find *t*-values are somewhat difficult to interpret, my figure does not show *t*-distributions, but the distribution of the effect size Cohen's d. This means that instead of having a critical *t*-value, we can read off the **critical d-value**. With n = 86 in each group the critical d-value is 0.3. This means that only effects larger than 0.3 will yield a $p < \alpha$. You can see the vertical line at the critical d-value.

Let's explore what this means in an interactive version of the figure above. Go to http://shiny.ieis.tue.nl/d_p_power/ where you will be able to change the values such as the sample size, the effect size, and the alpha level, that generate the figure. The default values for the shiny app are a true effect of d = 0.5, 50 participants per group, and an alpha level of 0.05. If you reload the webpage, it will reset to these default values.

The blue area in the figure is the **Type 2 error rate** (the probability of not finding $p < \alpha$, when there is a true effect), or **1 - power**. If the true effect size is d = 0.5, most of the observed effects will fall on the right of the critical d-value. To be precise, the statistical power based on an alpha of 0.05, 50 participants in each group, and assuming the true effect size is d = 0.5 is 69.69% (this number is provided in the text below the sliders). Thus, 30.31% of the effect sizes we will observe fall below d = 0.4 (our critical d-value) and will be non-significant.

Move the slider for Cohen's d to 0.79. You will now see that the blue area under the alternative distribution has the same size as the red area in the right tail of the null distribution. This distribution on the left is centered at 0 – it is the distribution of effect sizes that we can expect if the null hypothesis is true, and the effect size is exactly 0. The red areas in this distribution visualize Type 1 errors in a two-sided test. If the null-hypothesis is true, observing an effect larger than the critical d value is a Type 1 error in the positive direction. Because the figure illustrates a two-sided test, an equally extreme effect size in the negative direction (d > -0.4) would also be a Type 1 error.

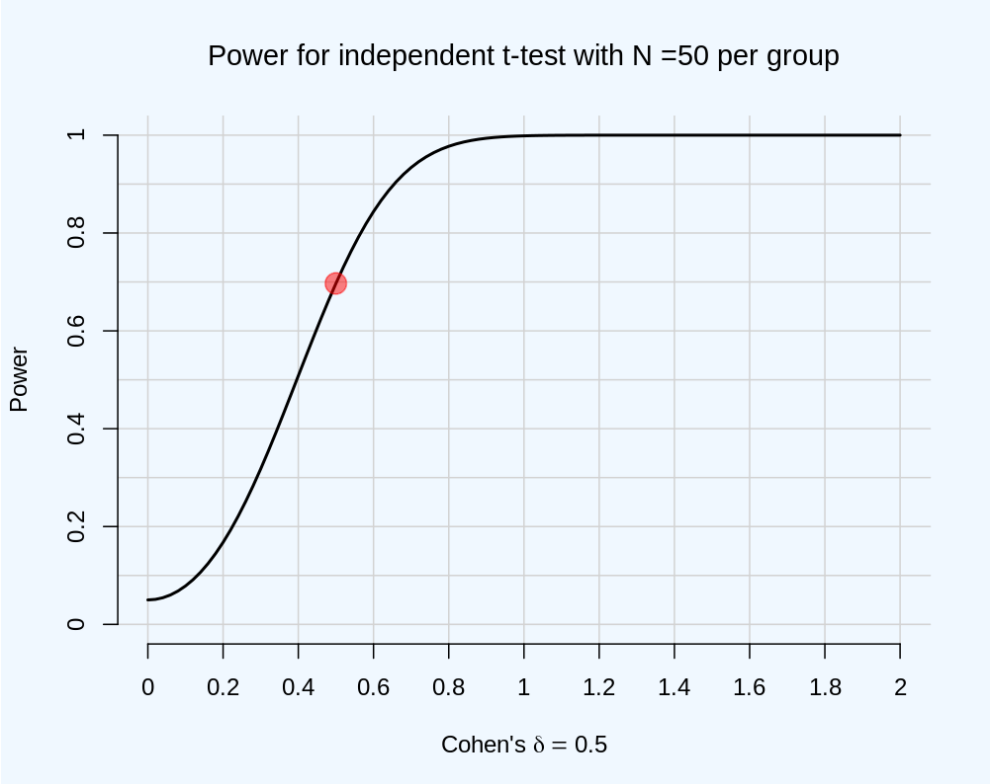
Q1: When the true d = 0.79 and n = 50 per group, the blue area is approximately the same size as one of the red areas, which means (feel free to use information in the text beneath the sliders):

A) The Type 1 error rate is approximately as large as the Type 2 error rate
B) The Type 1 error rate is approximately twice as large as the Type 2 error rate
C) The Type 2 error rate is approximately twice as large as the Type 1 error rate
D) The Type 1 error rate and the Type 2 error rate cannot be directly compared.

Q2: Three sliders influence what the figure looks like: The sample size per condition, the true effect size, and the alpha level. Which statement is true?

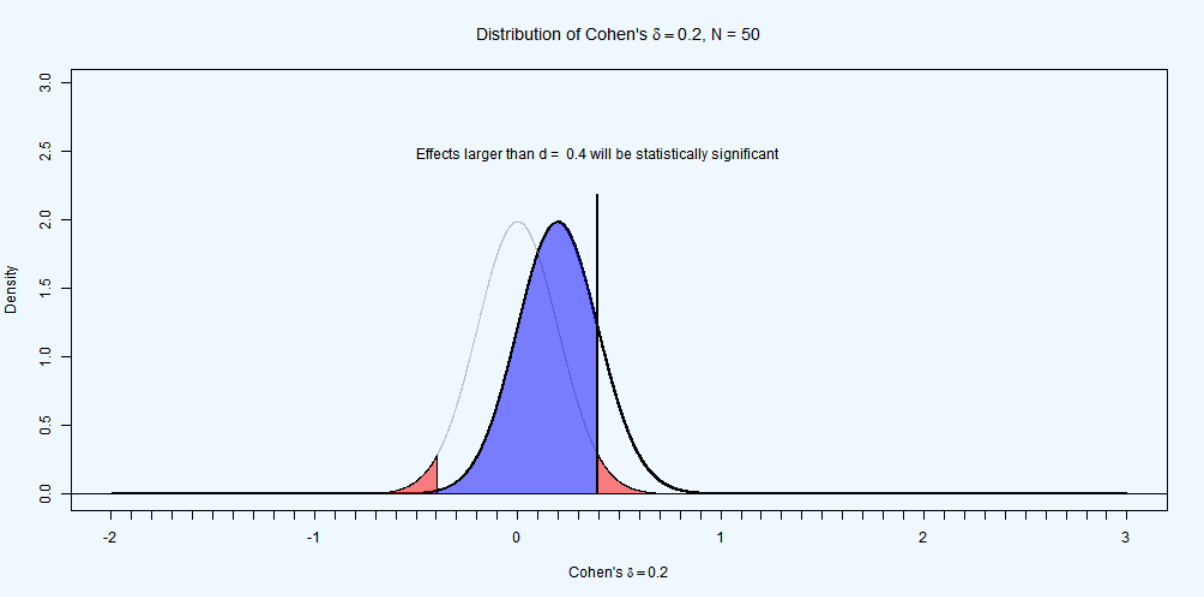
A) The critical d-value is influenced by the sample size per group, the true effect size, but **not** by the alpha level.
B) The critical d-value is influenced by the sample size per group, the alpha level, but **not** by the true effect size.
C) The critical d-value is influenced by the alpha level, the true effect size, but **not** by the sample size per group.
D) The critical d-value is influenced by the sample size per group, the alpha level, and by the true effect size.

Although it is common to talk about 'the power' of a study, statistical power is not a single value: It is a curve. When a study is designed, the alpha level and the sample size are chosen before the data is collected, and can thus be considered fixed when calculating power. But the true effect size is unknown, and therefore it makes more sense to think about power as a **power curve** across all possible effect sizes, as illustrated in the figure below.



This figure (which is also available in the online Shiny app, in the bottom right) visualizes that power is low for small effect sizes close to 0, being low for effects up to d = 0.4, and generally considered sufficient (upwards of 80%) from d = 0.6. Therefore, a study never has high power, or low power. It has high power for specific effects, and low power for other effects. When the true effect size is 0, power is formally undefined, but in the power formula, it ends up at the Type 1 error rate (which you can check in the shiny app by moving the alpha level up or down).

It needs a little explanation how we can have low power for small effects (e.g., in the figure above, 17% power for a **true effect** of d = 0.2), while the critical d-value is 0.4, and **observed effects** smaller than d = 0.4 will never be statistically significant. Both these statements are true, and the difference between a 'true' population effect and an 'observed' sample effect is very important. To see why, move the slider of Cohen's d to d = 0.2.



We see that the critical d-value has not changed: It is still true that only effects larger than d = 0.4 will be statistically significant. However, we also see from the distribution that we can expect some *observed effect sizes* to be larger than 0.4 when the *true population effect size* is d = 0.2 – the text tells us that 16.77% of *observed effect sizes* will be larger than 0.4 in the long run. That is not a lot, but it is something. For example, the critical d-value is 0.4. The true effect size is 0.2 in the *population*. In your *sample*, you will in the long run sometimes observe effect sizes larger than d = 0.4, and these effects would be statistically significant. In the figure above, if we have extreme publication bias and only significant results are published, we would end up with effect size estimates that are all larger than d = 0.4, even though the true effect size is 0.2.

It is now (hopefully) clear that it is possible to have low power for a *true effect size* that is smaller than the critical d-value. Even when the true effect size is smaller than the critical

d-value, some of the *observed effect sizes* of your samples will be larger than the critical d-value. However, only overestimated effect sizes (i.e., observed effect sizes that are larger than the true effect size) will be statistically significant. This is the reason why publication bias combined with underpowered research is problematic: It leads to a large **overestimation of the true effect size** when only observed effect sizes from statistically significant findings in underpowered studies end up in the scientific literature.

In addition to some basic insights into statistical power, the Shiny app can be used to see what the smallest observed effect size is that can be detected in a study with a **specific sample size** and **alpha level**. Because this **critical d-value** is independent of the true effect size (which is typically unknown) it is a way to evaluate the observed effect sizes that could have been statistically significant in a study with a certain sample size and alpha level.

Q3: Imagine researchers performed a study with 18 participants in each condition, and performed a t-test using an alpha level of 0.01. What is the smallest effect size that could have been statistically significant in this study?

A) d = 0.47
B) d = 0.56
C) d = 0.91
D) d = 1

Q4: You expect the true effect size in your next study to be d = 0.5, and you plan to use an alpha level of 0.05. You collect 30 participants in each group for an independent *t*-test. Which statement is true?

A) You have low power for all possible effect sizes.
B) You have sufficient (i.e., > 80%) power for all effect sizes you are interested in.
C) Observed effect sizes of d = 0.5 will never be statistically significant.
D) Observed effect sizes of d = 0.5 will be statistically significant.

Setting the smallest effect size of interest in replication studies

If you attempt to replicate a study, one justifiable option when choosing the smallest effect size of interest (SESOI) is to use the smallest observed effect size that could have been statistically significant in the study you are replicating. In other words, *you decide that effects that could not have yielded a p-value less than α in an original study will not be considered meaningful in the replication study*. **The assumption here is that the original authors were interested in observing a significant effect, and thus were not interested in observed effect sizes that could not have yielded a significant result**. It might be likely that the original authors did not consider which effect sizes their study had good statistical power to detect, or that they were interested in smaller effects

but gambled on observing an especially large effect in the sample purely as a result of random variation. Even then, when building on earlier research that does not specify a SESOI, a justifiable starting point might be to set the SESOI to the smallest effect size that, when observed in the original study, **would have been statistically significant**. Not all researchers might agree with this (e.g., the original authors might say they actually cared just as much about an effect of d = 0.001). However, as we try to change the field from the current situation where no one specifies what would falsify their hypothesis, or what their smallest effect size of interest is, this approach is one way to get started. I see this approach as a tennis match. Original authors serve and hit the ball across the net, saying 'look, something is going on'. The approach to set the SESOI to the effect size that could have been significant in the original study is a return volley which allows you to say 'there does not seem to be anything large enough that could have been significant in your own original study'. This is never the end of the match – the original authors can attempt to return the ball with a more specific statement about effects their theory predicts, and demonstrate such a smaller effect size is present.

In practice, this will often mean setting the SESOI to the effect size the original study had approximately 50% power to detect (given the fact that in large studies where the main test is an independent *t*-test, a study with $p = 0.05$ has an observed power of 50%, see Lenth, 2007). This effect size can be calculated using a sensitivity power analysis (for example in G*Power). In this sense, this approach is very similar to the small telescopes approach by Simonsohn (2015), except we calculate the effect size the original study had 50% power for, instead of 33% power.

Setting the SESOI based on theoretical predictions

If you are not building on previous studies, the SESOI can be set based on **theoretical predictions**, or a **cost-benefit analysis** that specifies the smallest effect size that would be practically relevant. Sometimes, an intervention should have a specific effect size to be worthwhile (for example, financially). Testing against a SESOI based on cost-benefit analysis allows you to decide if an intervention is worth it, when weighed against the costs.

Q5: Take a moment to think about whether the **theory** you are building on makes a quantifiable prediction that would allow you to set a smallest effect size of interest.

Q6: Take a moment to think about whether the topic you study allows you to make a **cost-benefit analysis**, where an effect size needs to be large enough to be worthwhile.

Setting the smallest effect size of interest based on resources

Not all experiments are designed to test quantifiable theoretical predictions, nor do they allow for a cost-benefit analysis. Researchers often have more precise ideas about the

amount of data that they can afford to collect, or that **other researchers in their field commonly collect**, than about the effect size they predict or that would be worth studying. The amount of data that is collected limits the inferences one can make. Given the alpha level and the planned sample size for a study, researchers can **calculate the smallest effect size that they have sufficient resources to detect**.

Setting the smallest effect size of interest based on this approach does not answer any theoretical question (after all, the SESOI is not based on any theoretical prediction). Instead, it answers a **resource question**: Given an available sample size, or given a sample size common in a field, is the effect large enough so that we can reliably study it? If the answer is 'NO' **this does not mean that the effect is not interesting per se** – it is just not interesting given the resources researchers have available. If effects larger than a resource based SESOI are rejected, a field might decide that it is time to examine the research question collaboratively, by coordinating research lines, and collecting enough data in a team to reliably study it.

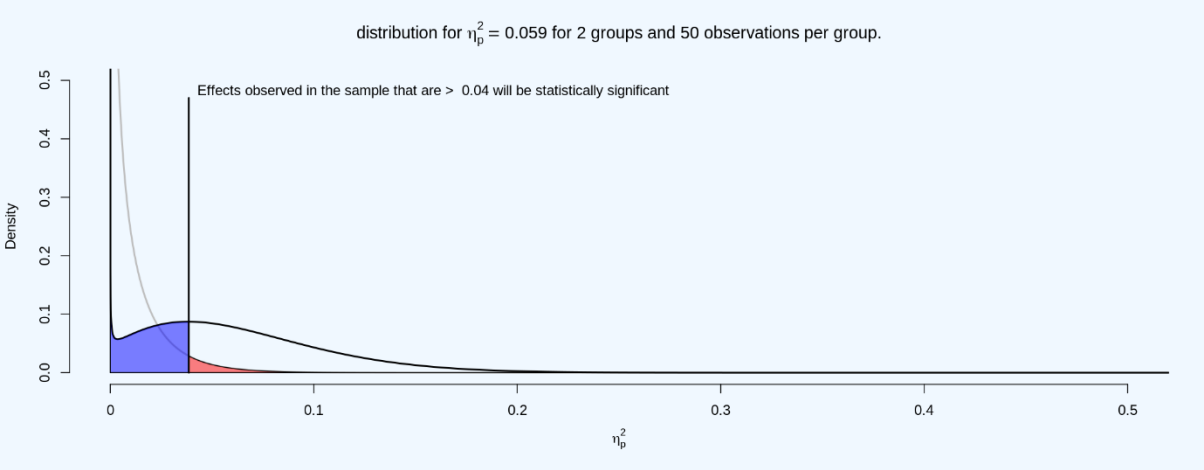
For example, imagine a line of research in which a hypothesis has almost always been tested by performing a one-sample t-test on a sample size smaller than 100 observations. A one-sample t test on 100 observations, using an alpha of .05 (two sided), has 80% power to detect an effect of d = 0.28 (as can be calculated in power software such as G*Power, or using the R code below).

```
library("pwr")
pwr.t.test(n = 100, sig.level = 0.05, power = 0.80, type = "one.sample",
alternative = "two.sided")$d
```

In a new study, concluding that one can reliably reject the presence of effects more extreme than d = 0.28 suggests that sample sizes of 100 might not be enough to detect effects in such research lines. Rejecting the presence of effects more extreme than d = 0.28 does not test a theoretical prediction, but it contributes to the literature by answering a **resource question**. It suggests that future studies need to collect data on samples larger than 100 observations to examine this hypothesis. Note that statisticians and theoreticians are likely not impressed by such resource questions, but they are, I think, very relevant for researchers who have to decide whether they can actually pursue a line of research in practice in individual labs given their resources. If smaller effects are deemed worthwhile to study, and each lab can realistically collect 100 observations per study, **it is time to collaborate**. An example of such a collaboration is the Psychological Science Accelerator.

The example we have used above concerned an independent *t*-test, but the idea can be generalized. A shiny app for an F-test is available here: http://shiny.ieis.tue.nl/f_p_power/. The effect size associated to the power of an F-test is partial eta squared (η_p^2), which for a One-Way ANOVA (visualized in the Shiny app) equals eta-squared.

The distribution for eta-squared looks slightly different from the distribution of Cohen's d, primarily because an *F*-test is a one-directional test (and because of this, eta-squared values are all positive, while Cohen's d can be positive or negative). The light grey line plots the expected distribution of eta-squared when the null is true, with the red area under the curve indicating Type 1 errors, and the black line plots the expected distribution of eta-squared when the true effect size is η = 0.059. The blue area indicates the expected effect sizes smaller that the critical η of 0.04, which will not be statistically significant, and thus will be Type 2 errors.



Q7: Set the number of participants (per condition) to 14, and the number of groups to 3. Which effect sizes (expressed in partial eta-squared, as indicated on the vertical axis) can be statistically significant with n = 14 per group, and 3 groups?

A) Only effects larger than 0.11
B) Only effects larger than 0.13
C) Only effects larger than 0.14
D) Only effects larger than 0.16

Every sample size and alpha level implies observed effect sizes that can be statistically significant in your study. Looking at which observed effects you can detect is a useful way to make sure you could actually detect the smallest effect size you are interested in.