Setting the Smallest Effect Size of Interest in Replications Studies Using the Small Telescopes Approach

Ideally, as De Groot (1969) mentioned, researchers who publish novel research would always specify which effects would disprove their theory. Regrettably, this is not yet common practice. This is particularly problematic when a researcher performs a close replication of earlier work. Because it is never possible to prove an effect is exactly zero, and the original authors seldom specify which range of effect sizes would falsify their hypotheses, it has proven to be very difficult to interpret the outcome of a replication study. When does the new data contradict the original finding?

Consider a study in which you want to test the idea of the wisdom of crowds. You ask 20 people to estimate the number of coins in a jar, expecting the average to be very close to the true value. The research question is whether the people can on average correctly guess the number of coins, which is 500. The observed mean guess by 20 people is 550, with a standard deviation of 100. The observed difference from the true value is statistically significant, $t(19)=2.37$, $p = 0.0375$, with a Cohen’s d of 0.5. Can it really be that the group average is so far off? Is there no Wisdom of Crowds? Was there something special about the coins you used that make it especially difficult to guess their number? Or was it just a fluke? You set out to perform a close replication of this study.

You want your study to be informative, regardless of whether there is an effect or not. This means you need to design a replication study that will allow you to draw an informative conclusion, regardless of whether the alternative hypothesis is true (the crowd will not estimate the true number of coins accurately) or whether the null hypothesis is true (the crowd will guess 500 coins, and the original study was a fluke). But since the original researcher did not specify a smallest effect size of interest, when would a replication study allow you to conclude the original study is contradicted by the new data? Observing a mean of exactly 500 would perhaps by some be considered quite convincing, but due to random variation you will (almost) never find a mean score of exactly 500. A non-significant result can’t be interpreted as the absence of an effect, because your study might have too small a sample size to detect meaningful effects. So how can we move forward and define an effect size that is meaningful? How can you design a study that has the ability to disconfirm a previous finding?

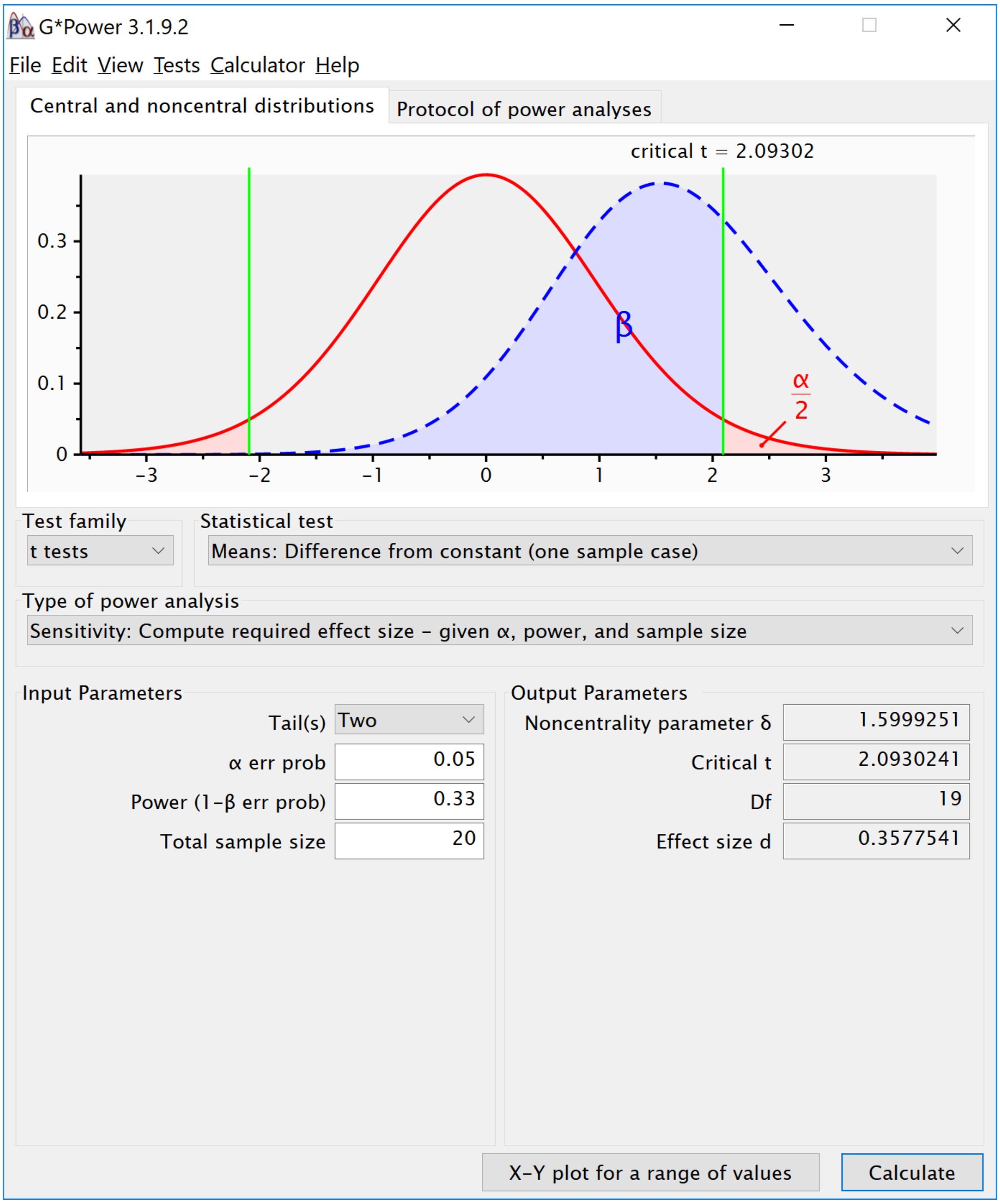
Uri Simonsohn (2015) defines a small effect as "**one that would give 33% power to the original study**". In other words, the effect size that would give the original study odds of 2:1 *against* observing a statistically significant result if there was an effect. The idea is that if the original study had 33% power, the probability of observing a significant effect, if there was a true effect, is too low to reliably distinguish signal from noise (or situations where there is a true effect from situations where there is no true effect). Simonsohn (2015, p. 561) calls this the **small telescopes approach**, and writes: "Imagine an astronomer claiming to have found a new planet with a telescope. Another astronomer tries to replicate the discovery using a larger telescope and finds nothing. Although this does not prove that the planet does not exist, it does nevertheless contradict the original findings, because planets that are observable with the smaller telescope should also be observable with the larger one."

Although this approach to setting a smallest effect size of interest (SESOI) is arbitrary (why not 30% power, or 35%?) it suffices for practical purposes (and you are free to choose a power level you think is too low). The nice thing about this definition of a SESOI is that **if you know the sample size of the original study, you can always calculate the effect size that study had 33% power to detect**. You can thus always use this approach to set a smallest effect size of interest. If you fail to find support for an effect size the original study has 33% power to detect, it does not mean there is no true effect, and not even that the effect is too small to be of any theoretical or practical interest. But using the small telescopes approach is a good **first step**, since it will get the conversation started about which effects are meaningful and allows researchers who want to replicate a study to specify when they would consider the original claim falsified.

With the small telescopes approach, the SESOI is based **only on the sample size in the original study**. A smallest effect size of interest is set only for **effects in the same direction**. All effects smaller than this effect (including large effects in the opposite direction) are interpreted as a failure to replicate the original results. This makes the small telescopes approach formally an inferiority test, where we try to reject the hypothesis that the effect is as large or larger than the effect the original study has 33% power to detect. In this sense, it is a simple one-sided test, not against 0, but against a SESOI.

For example, consider our study above in which 20 guessers tried to estimate the number of coins. The results were analyzed with a two-sided one-sample *t*-test, using an alpha level of 0.05. To determine the effect size that this study had 33% power for, we can perform a sensitivity analysis. In a sensitivity analysis we compute the required effect size given the alpha, sample size, and desired statistical power. Note that Simonsohn uses a two-sided test in his power analyses, which we will follow here – if the original study reported a pre-registered directional prediction, the power analysis should be based on a one-sided test. In this case, the alpha level is 0.05, the total sample size is 20, and the desired power is 33%. We compute the effect size that gives us 33% power and see that it is a Cohen's d of 0.358. This means we can set our smallest effect size of interest for the replication study to *d* = 0.358. If we can reject effects as large or larger than *d* = 0.358, we can conclude that the effect is smaller than anything the original study had 33% power for. The screenshot below illustrates the correct settings in G*power, and the code in R is:

```
library("pwr")
pwr.t.test(n = 20, sig.level = 0.05, power = 0.33, type = "one.sample",
alternative = "two.sided")
```


G*Power 3.1.9.2
File Edit View Tests Calculator Help
Central and noncentral distributions
Protocol of power analyses
critical t = 2.09302
0.3
0.2
0.1
0
-3
-2
-1
0
1
2
3
β
α
2
X-Y plot for a range of values
Calculate

Test family: t tests

Statistical test: Means: Difference from constant (one sample case)

Type of power analysis: Sensitivity: Compute required effect size – given α, power, and sample size

Input Parameters

Tail(s): Two
α err prob: 0.05
Power (1-β err prob): 0.33
Total sample size: 20

Output Parameters

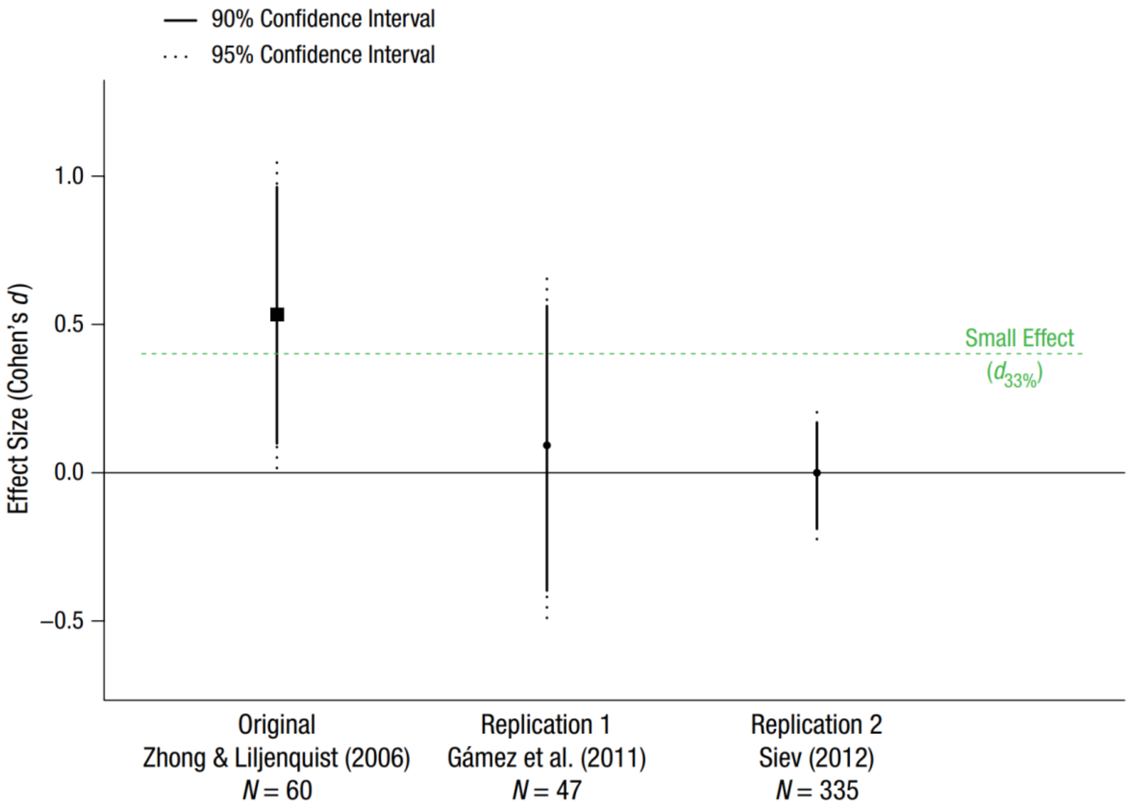
Noncentrality parameter δ: 1.5999251
Critical t: 2.0930241
Df: 19
Effect size d: 0.3577541

Determining the SESOI based on the effect size the original study had 33% power to detect has an additional convenient property. Imagine the true effect size is actually 0, and you perform a statistical test to see if the data is statistically smaller than the SESOI based on the small telescopes approach (which is called an inferiority test). If you increase the sample size by 2.5 times, you will have approximately 80% power for this inferiority test, assuming the true effect size is exactly 0 (e.g., d = 0). People who do a replication study can follow the small telescope recommendations, and very easily determine both the **smallest effect size of interest**, and the **sample size** needed to design an informative replication study.

The figure below, from Simonsohn (2015) illustrates the small telescopes approach using a real-life example. The original study by Zhong and Liljenquist (2006) had a tiny sample size of 30 participants in each condition and observed an effect size of d = 0.53, which was barely statistically different from zero. Given a sample size of 30 per condition, the study had 33% power to detect effects larger than d = 0.401. This "small effect" is indicated by the green dashed line. In R, the smallest effect size of interest is calculated using:

```
library("pwr")
pwr.t.test(n = 30, sig.level = 0.05, power = 1/3, type = "two.sample",
alternative = "two.sided")
```

Note that 33% power is a rounded value, and the calculation uses 1/3 (or 0.3333333...).



We can see that the first replication by Gámez and colleagues also had a relatively small sample size (N = 47, compared to N = 60 in the original study), and was not designed to yield informative results when interpreted with a small telescopes approach. The confidence interval is very wide and includes the null effect (d = 0) and the smallest effect size of interest (d = 0.401). Thus, this study is inconclusive. We can't reject the null, but we can also not reject effect sizes of 0.401 or larger that are still considered to be in line with the original result. The second replication has a much larger sample size, and tells us that we can't reject the null, but we can reject the smallest effect size of interest, suggesting that the effect is smaller than what is considered an interesting effect based on the small telescopes approach.

Although the *small telescope* recommendations are easy to use, one should take care not to turn any statistical procedure into a heuristic. In our example above with the 20 referees, a Cohen's d of 0.358 would be used as a smallest effect size of interest, and a sample size of 50 would be collected (2.5 times the original 20), but if someone would make the effort to perform a replication study, it would be relatively easy to collect a larger sample size. Alternatively, had the original study been extremely large, it would have had high power for effects that might not be practically significant, and we would not want to collect 2.5 times as many observations in a replication study. Indeed, as Simonsohn writes: "whether we need 2.5 times the original sample size or not depends on the question we wish to answer. If we are interested in testing whether the effect size is smaller than d33%, then, yes, we need about 2.5 times the original sample size no matter how big that original sample was. When samples are very large, however, that may not be the question of interest." This nicely fits with the main theme of this course: Always think about the question you want to ask! Do not automatically follow a 2.5 times n heuristic, and always reflect on whether the use of a suggested procedure is appropriate in a given situation.

Q1. In the example above we calculated the SESOI for a one-sample *t*-test. Adjust the R code (or use G*power) to calculate power for a **two-sample *t*-test** with an **alpha level of 0.05**, **n = 20** in each condition (you only need to change "one.sample" into "two.sample" in the R code, or choose the correct option in G*power). What is the SESOI based on the small telescope approach? Note that for this answer, it happens to depend on whether you enter the power as 0.33 or 1/3 (or 0.333).

A) d = 0.25 (setting power to 0.33) or 0.26 (setting power to 1/3)
B) d = 0.33 (setting power to 0.33) or 0.34 (setting power to 1/3)
C) d = 0.49 (setting power to 0.33) or 0.50 (setting power to 1/3)
D) d = 0.71 (setting power to 0.33) or 0.72 (setting power to 1/3)

Q2. Let's assume you are trying to replicate a previous result based on a correlation in a two-sided test. The study had 150 participants. Calculate the SESOI for a replication of this study that will use an alpha level of 0.05, based on the small telescopes approach using either G*Power or R. Note that for this answer, it happens to depend on whether you enter the power as 0.33 or 1/3 (or 0.333). In R, you need the code:

```
pwr.r.test(n = X, sig.level = X, power = X, alternative = "two.sided")
```

A) r = 0.124 (setting power to 0.33) or 0.125 (setting power to 1/3)
B) r = 0.224 (setting power to 0.33) or 0.225 (setting power to 1/3)
C) r = 0.226 (setting power to 0.33) or 0.227 (setting power to 1/3)
D) r = 0.402 (setting power to 0.33) or 0.403 (setting power to 1/3)

Q3. In the age of big data researchers often have access to large databases, and can run correlations on samples of thousands of observations. Let's assume the original study in the previous question did not have 150 observations, but 15000 observations. We still use an alpha level of 0.05. Note that for this answer, it happens to depend on whether you enter the power as 0.33 or 1/3 (or 0.333). What is the SESOI based on the small telescopes approach? And is this effect likely to be practically or theoretically significant?

A) r = 0.0124 (setting power to 0.33) or 0.0125 (setting power to 1/3)
B) r = 0.0224 (setting power to 0.33) or 0.0225 (setting power to 1/3)
C) r = 0.0226 (setting power to 0.33) or 0.0227 (setting power to 1/3)
D) r = 0.0402 (setting power to 0.33) or 0.0403 (setting power to 1/3)

Q4: Using the small telescopes approach, you set the SESOI in a replication study to d = 0.35, and set the alpha level to 0.05. After collecting the data in a well-powered replication study that was as close to the original study as practically possible, you find no significant effect, and you can reject effects as large or larger than d = 0.35. What is the correct interpretation of this result?

A) There is no effect.
B) We can statistically reject (using an alpha of 0.05) effects anyone would find theoretically meaningful.
C) We can statistically reject (using an alpha of 0.05) effects anyone would find practically relevant.
D) We can statistically reject (using an alpha of 0.05) effects as large or larger than 0.35.

References:

de Groot, A. D. (1969). Methodology (Vol. 6). The Hague: Mouton & Co.

Simonsohn, U. (2015). Small Telescopes Detectability and the Evaluation of Replication Results. Psychological Science, 26(5), 559–569. https://doi.org/10.1177/0956797614567341