# Effect Sizes

What is the most important outcome of an empirical study? You might be tempted to say it's the *p*-value of the statistical test, given that it is practically always reported in articles, and determines whether we call something 'significant' or not. However, as the famous statistician Cohen writes in his 1990 article 'Some Things I've Learned (So Far)': "I have learned and taught that the primary product of a research inquiry is one or more measures of effect size, not *p*-values".

A measure of effect size is a quantitative description of the strength of a phenomenon. It is expressed as a number on a scale, and which scale is used depends on the effect size measure that is used. For **unstandardized effect sizes**, we can use a scale that people are very familiar with. For example, children grow on average 6 centimeters a year between the age of 2 and puberty. We can interpret 6 centimeters a year as an effect size. It is obvious an effect size has many benefits over a *p*-value. A *p*-value gives an indication that it is very unlikely children stay the same size as they become older – effect sizes tell us what size clothes we can expect children to wear when they are a certain age, and how long it will take before their new clothes are too small.

Researchers often report **standardized effect sizes** because many psychological variables are not measured on a scale people are familiar with, or are often measured on different scales. If you ask people how happy they are, an answer of '5' will mean something very different if you asked then on a scale from 1 to 5 than if you asked them on a scale from 1 to 9. Standardized effect sizes allow researchers to present the magnitude of the reported effects in a standardized metric. Therefore, standardized effect sizes can be understood and compared regardless of the scale that was used to measure the dependent variable. Such standardized effect sizes allow researchers to communicate the practical significance of their results (the practical consequences of the findings for daily life), instead of only reporting the statistical significance (how surprising is the data, given the assumption that there is no effect in the population).

Standardized effect sizes also allow researchers to draw **meta-analytic conclusions** by comparing standardized effect sizes across studies. In a meta-analysis, researchers look at the results of a large number of studies and calculate the average effect size across studies to draw more reliable conclusions. Finally, standardized effect sizes from previous studies can be used when planning a new study. An **a-priori power analysis** can provide an indication of the average sample size a study needs to observe a statistically significant result with a desired probability.

It is important to make a distinction between '**statistically significant**' and '**substantially interesting**'. For example, we might be able to reliably measure that on average, men who are 19 years old will grow another 20 millimeters before they are 21. This difference might very well be statistically significant, but if you go shopping for clothes when you are a 19-year old man, it is not something you need to think about. The most important way to evaluate whether an effect is substantially interesting is to look at the effect size.

## The Facebook experiment

In the summer of 2014 there were some concerns about an experiment Facebook had performed on its users to examine 'emotional mood contagion', or the idea that people's moods can be influenced by the mood of people around them. You can read the article here. For starters, there was substantial concern about the ethical aspects of the study, primarily because the researchers who performed the study had not asked **informed consent** from the participants in the study (you and me), nor did they ask for permission from the **institutional review board** (or ethics committee) of their university.

One of the other criticisms on the study was that it could be dangerous to influence people's mood. As Nancy J. Smyth, dean of the University of Buffalo's School of Social Work wrote on her Social Work blog: "There might even have been increased self-harm episodes, out of control anger, or dare I say it, suicide attempts or suicides that resulted from the experimental manipulation. Did this experiment create harm? The problem is, we will never know, because the protections for human subjects were never put into place".

If this Facebook experiment had such a strong effect on people's mood that it made some people commit suicide who would otherwise not have committed suicide, this would obviously be problematic. So let us look at the effects the manipulation Facebook used had on people a bit more closely.

From the article, let's see what the researchers manipulated:

*Two parallel experiments were conducted for positive and negative emotion: One in which exposure to friends' positive emotional content in their News Feed was reduced, and one in which exposure to negative emotional content in their News Feed was*

*reduced. In these conditions, when a person loaded their News Feed, posts that contained emotional content of the relevant emotional valence, each emotional post had between a 10% and 90% chance (based on their User ID) of being omitted from their News Feed for that specific viewing.*

Then what they measured:

*For each experiment, two dependent variables were examined pertaining to emotionality expressed in people's own status updates: the percentage of all words produced by a given person that was either positive or negative during the experimental period. In total, over 3 million posts were analyzed, containing over 122 million words, 4 million of which were positive (3.6%) and 1.8 million negative (1.6%).*

And then what they found:

*When positive posts were reduced in the News Feed, the percentage of positive words in people's status updates decreased by B = −0.1% compared with control [t(310,044) = −5.63, P < 0.001, Cohen's d = 0.02], whereas the percentage of words that were negative increased by B = 0.04% (t = 2.71, P = 0.007, d = 0.001). Conversely, when negative posts were reduced, the percent of words that were negative decreased by B = −0.07% [t(310,541) = −5.51, P < 0.001, d = 0.02] and the percentage of words that were positive, conversely, increased by B = 0.06% (t = 2.19, P < 0.003, d = 0.008).*

Here, we will focus on the negative effects of the Facebook study (so specifically, the increase in negative words people used) to get an idea of whether there is a risk of an increase in suicide rates. Even though apparently there was a negative effect, it is not easy to get an understanding about the size of the effect from the numbers as mentioned in the text. Moreover, the number of posts that the researchers analyzed was really large. With a large sample, it becomes important to check if the size of the effect is such that the finding is substantially interesting, because with large sample sizes even minute differences will turn out to be statistically significant (we will look at this in more detail below). For that, we need a better understanding of "effect sizes".

Now that we realize why effect sizes are important, let us look more closely at the most commonly used effect sizes, and how these are calculated.

Effect sizes can be grouped into two families (Rosenthal, 1994): The ***d* family** (based on standardized mean differences) and the ***r* family** (based on measures of strength of association). Conceptually, the *d* family effect sizes are based on a comparison between the difference between the observations, divided by the standard deviation of these observations. This means that a Cohen's *d* = 1 means the standardized difference between two groups equals one standard deviation. The *r* family effect sizes are based on the proportion of variance that is explained by group membership (e.g., a correlation of $r = 0.5$ indicates 25% ($r^2$) of the variance is explained by the difference between groups). Don't worry if you do not exactly get what this means at this point. The crucial issue is that we need to understand how to interpret the size of an effect, and that there are different ways to express the size of this effect.

## Cohen's *d*

The size of the effect in the Facebook study is given by the statistic Cohen's *d* (which we will discuss in more detail below). Cohen's *d* (the *d* is italicized) is used to describe the standardized mean difference of an effect. This value can be used to compare effects across studies, even when the dependent variables are measured in different ways, for example when one study uses 7-point scales to measure dependent variables, while the other study uses 9-point scales, or even when completely different measures are used, such as when one study uses self-report measures, and another study used physiological measurements.

Cohen's *d* ranges from 0 to infinity. Before we get into the statistical details, let's first visualize what a Cohen's d of 0.001 (as was found in the Facebook study) means.

Go to http://rpsychologist.com/d3/cohend/, a website that allows you to visualize the differences between two measurements (such as the increase in negative words used by the Facebook user when the number of positive words on the timeline was reduced). First, you need to click the wheel behind the 'Slide me' text: Set 'Max Cohen's *d*' to 0.002 and 'Slider step size' to 0.001 and press ENTER. As you might have guessed by now, the effect sizes in behavioral research are typically a lot larger, so we have to reduce the scale considerably. Now slide the slider to the middle so that we visualize Cohen's *d* = 0.001. Read the 'Common language explanation' below the figure.

Q1) Below the graph, you see some ways to interpret Cohen's d in non-mathematical terms (the summary is provided about a number of people, but in our case, we are

examining numbers of words). How many words (rounded to a whole number) does a person need to type before 1 word will be more negative (instead of positive)?

This illustrates the difference between a statistical difference and practical significance (or substantial interest). The effect is so small that it is unlikely to be noticeable for a single individual. Hence, in this case, and without further evidence, we would not worry too much about the extra suicides the research could have caused. Nevertheless, even such small effects can matter in other kinds of research. If an intervention makes people spend more money with a *d* = 0.001, and you have millions of transactions a year, a very small effect might very well make you a lot of money.

Change the settings back to 'Max Cohen's *d* to 2 and 'Slider step size' to 0.01. Large meta-analytic efforts by Richard, Bond, and Stookes-Zoota (2003) have estimated the average effect size in psychological studies to have a Cohen's d = 0.43. Let's try an example to get a feel for the size of such an effect.

Q2) Assume we know that people are more likely to comply with a large request after an initial smaller request, than when you ask the large request directly (this is known as the foot-in-the-door effect), and that in a specific context this effect size is 0.43. Given this effect size, how likely is it that a random person is drawn from the 'small initial request condition' will be more likely to agree with your larger request, compared to a person in the 'no initial small request' condition? (use the slider to answer this question, and round to a whole number – answer by typing in only the number without a %)

Q3) Now set the slider to a Cohen's *d* of 2. Based on this data, the difference between the height of 21-year old men and women in The Netherlands is approximately 13 centimeters (in an unstandardized effect size), or a standardized effect size of *d* = 2. If I pick a random man and a random woman walking down the street in my hometown of Rotterdam, how likely is it that the man will be taller than the woman (round the percentage to a whole number – answer by typing in only the number without a %)?

To understand Cohen's *d*, let's first look at the formula for the *t*-statistic:

$$t = \frac{\bar{M}_1 - \bar{M}_2}{SD_{pooled} \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Here $\bar{M}_1 - \bar{M}_2$ is the difference between the means, and $SD_{pooled}$ is the pooled standard deviation (see Lakens, 2013), and $n_1$ and $n_2$ are the sample sizes of the two groups you are comparing. The *t*-value (because it follows a known distribution) is used to determine whether the difference between two groups in a *t*-test is statistically significant. The formula for Cohen's *d* is very similar:

$$\text{Cohen's } d = \frac{\bar{M}_1 - \bar{M}_2}{SD_{pooled}}$$

You can calculate Cohen's *d* by hand from the independent samples *t*-value (which can often be convenient when the result section of the paper you are reading does not report effect sizes) through:

$$\text{Cohen's } d = t \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

As you can see, the sample size is part of the formula for a *t*-value, but it is not part of the formula for Cohen's *d*.

Q4) Let's assume the difference between two means we observe is 1, and the pooled standard deviation is also 1. What, on average, happens to the *t*-value and Cohen's *d*, as we would simulate studies, as a function of the sample size in these simulations?

A) Given the mean difference and standard deviation, as the sample size becomes bigger, the *t*-value become larger, and Cohen's *d* becomes larger.

B) Given the mean difference and standard deviation, as the sample size becomes bigger, the *t*-value gets closer to the true value, and Cohen's *d* becomes larger.

C) Given the mean difference and standard deviation, as the sample size becomes bigger, the *t*-value become larger, and Cohen's *d* gets closer to the true value.

D) Given the mean difference and standard deviation, as the sample size becomes bigger, the *t*-value gets closer to the true value, and Cohen's *d* gets closer to the true value.

## Effect sizes differ depending on whether they correct for bias or not

Population effect sizes are almost always estimated on the basis of samples, and as a measure of the population effect size estimate based on sample averages, Cohen's *d* overestimates the true population effect (when Cohen's *d* refers to the population, the Greek letter δ is often used). Therefore, corrections for bias are used (even though these corrections do not always lead to a completely unbiased effect size estimate). In the *d* family of effect sizes, the correction for bias in the population effect size estimate of Cohen's δ is known as Hedges' *g* (although different people use different names – $d_{unbiased}$

is also used). This correction for bias is only really noticeable in small sample sizes, but since we often use software to calculate effect sizes anyway, it makes sense to always report Hedge's *g* instead of Cohen's *d*.

A commonly used interpretation of Cohen's *d* is to refer to effect sizes as small (*d*=0.2), medium (*d*=0.5), and large (*d*=0.8) based on benchmarks suggested by Cohen (1988) – note, in the video I talk about d = 0.3 being a small effect size, but 0.2 is the benchmark for a small effect as specified by Cohen). However, these values are arbitrary and should not be interpreted too rigidly. Furthermore, small effect sizes can have large consequences, such as an intervention that leads to a reliable reduction in suicide rates with an effect size of *d*=0.1. On the other hand, you have to start somewhere in getting a feeling for effect sizes, and these benchmarks are a good starting point.

An interesting, though not often used, interpretation of differences between groups can be provided by the common language effect size, also known as the probability of superiority. It expresses the probability that a randomly sampled person from one group will have a higher observed measurement than a randomly sampled person from the other group (for between designs) or the other measurement (for within-designs) the probability that an individual has a higher value on one measurement than the other. We used it earlier and it is provided by the website that visualizes Cohen's d.

## *r* (correlations)

The second effect size that is widely used is *r*. You might remember that *r* is used to refer to a correlation. The correlation of two continuous variables can range from 0 (completely unrelated) to 1 (perfect positive relationship) or -1 (perfect negative relationship). Obviously, given the flexibility of human behavior (free will has a lot to do with it) correlations between psychological variables are rarely 1. The average effect size *r* in psychology is .21. As mentioned earlier, the *r* family effect sizes describe the proportion of variance that is explained by the independent variable, or $r^2$.

Earlier, I mentioned the average effect size in psychology is *d* = 0.43. You might, therefore, think a *d* = 0.43 and an *r* = .21 should be related somehow, and they are:

$$r = \frac{d_s}{\sqrt{d_s^2 + \frac{N^2 - 2N}{n_1 \times n_2}}}$$

The subscript s underneath Cohen's *d* is used to specify this Cohen's *d* is calculated based on the sample, not based on the population. This is almost always the case (except in simulation studies, where you can set the effect size in the population), and *N* is the total sample size of both groups, whereas $n_1$ and $n_2$ are the sample sizes of the two groups you are comparing.

Go to http://rpsychologist.com/d3/correlation/ to look at a good visualization of the proportion of variance that is explained by group membership, and the relationship between *r* and $r^2$. Look at the scatterplot and the shared variance for an effect size of *r* = .21 (Richard, Bond, & Stookes-Zoota, 2003).

Q5) Given that *r* = 0.21 was the average effect size in psychological research, how much variance in the data do we on average explain?

A) 2.1%

B) 21%

C) 4.4%

D) 44%

Q6) By default, the sample size on the website is 100. Change it to 500. What happens?

A) The proportion of explained variance is 5 times as large.

B) The proportion of explained variance is 5 times as small.

C) The proportion of explained variance is $5^2$ times as large.

D) The proportion of explained variance stays the same.

Feel free to play around with the visualization. At the very least, see what happens when the correlation is exactly 0, and exactly 1.

Effect sizes can be implausibly large. Let's take a look at a study that actually examines the number of suicides – as a function of the amount of country music played on the radio. You can find the paper here (for a free PDF version, click here). It won an Ig Nobel prize for studies that first make you laugh, and then think. I guess in this case the study should make you think about the importance of interpreting effect sizes.

The authors predicted the following:

*"We contend that the themes found in country music-foster a suicidal mood among people already at risk of suicide and that it is thereby associated with a high suicide rate."*

Then they collected data:

*"Our sample is comprised of 49 large metropolitan areas for which data on music were available. Exposure to country music is measured as the proportion of radio airtime devoted to country music. Suicide data were extracted from the annual Mortality Tapes, obtained from the Inter-University Consortium for Political and Social Research (ICPSR) at the University of Michigan. The dependent variable is the number of suicides per 100,000 population."*

And they concluded:

*"A significant zero-order correlation was found between white suicide rates and country music ($r = .54$, $p < .05$). The greater the airtime given to country music, the greater the white suicide rate."*

Cohen (1988) has provided benchmarks to define small ($r = 0.1$), medium ($r = 0.3$), and large ($r = 0.5$) effects. This means the effect of listening to country music on suicide rates is large. Remember that it is preferable to relate the effect size to other effects in the literature instead of to these benchmarks.

Q7) What do you think of the likelihood that listening to country music is strongly associated with higher suicide rates? Is country music really that bad? Write down your thoughts in a single sentence.

If you were doubtful about the possibility that this effect was real, you might not be surprised by the fact that other researchers were not able to reproduce the analysis of the original authors. It is likely that the results are spurious, or a Type 1 error.

Eta squared η² (part of the *r* family of effect sizes, and an extension of r that can be used for more than two sets of observations) measures the proportion of the variation in Y that is associated with membership of the different groups defined by X, or the sum of squares of the effect divided by the total sum of squares:

$$\eta^2 = \frac{SS_{effect}}{SS_{total}}$$

An η² of .13 means that 13% of the total variance can be accounted for by group membership. Although η² is an efficient way to compare the sizes of effects within a study (given that every effect is interpreted in relation to the total variance, all η² from a single study sum to 100%), eta squared cannot easily be compared between studies, because the total variability in a study ($SS_{total}$) depends on the design of a study, and increases when additional variables are manipulated (e.g., when independent variables are added). Keppel (1991) has recommended partial eta squared ($\eta_p^2$) to improve the comparability of effect sizes between studies, which expresses the sum of squares of the effect in relation to the sum of squares of the effect and the sum of squares of the error associated with the effect. Partial eta squared is calculated as:

$$\eta_p^2 = \frac{SS_{effect}}{SS_{effect} + SS_{error}}$$

For designs with fixed factors (manipulated factors, or factors that exhaust all levels of the independent variable, such as alive vs. dead), but not for designs with measured factors or covariates, partial eta squared can be computed from the *F*-value and its degrees of freedom (e.g., Cohen, 1965):

$$\eta_p^2 = \frac{F \times df_{effect}}{F \times df_{effect} + df_{error}}$$

For example, for an $F(1, 38) = 7.21$, $\eta_p^2 = 7.21 \times 1/(7.21 \times 1 + 38) = 0.16$.

Eta squared can be transformed into Cohen's *d*:

$$d = 2 \times f \text{ where } f^2 = \eta^2/(1 - \eta^2)$$

As with Cohen's *d*, η² is a biased estimate of the true effect size in the population. Two less biased effect size estimates have been proposed, epsilon squared $\varepsilon^2$ and omega squared $\omega^2$. Because these effect sizes are less biased, it is always better to use them. Partial epsilon squared and partial omega squared can be calculated based on the *F*-value and degrees of freedom.

$$\omega_p^2 = \frac{F - 1}{F + \frac{df_{error} + 1}{df_{effect}}}$$

$$\varepsilon_p^2 = \frac{F - 1}{F + \frac{df_{error}}{df_{effect}}}$$

A spreadsheet document to calculate $\eta_p^2$ $\omega_p^2$ $\varepsilon_p^2$ from the *F*-value and degrees of freedom is available from https://osf.io/sjgv4/.

**Practical Examples**

Q8) Download the spreadsheet from https://osf.io/sjgv4/. For a result of $F(1,38) = 6.3$, when all variables are manipulated, what is the partial epsilon squared effect size?

A) 0.142

B) 0.117

C) 0.120

For some unexplainable reason, widely used software such as SPSS did not bother to provide Cohen's *d* or Hedges' *g* in the default output for a *t*-test. I've created an easy to use a spreadsheet that will allow you to calculate effect sizes for within and between participant t-tests. You can download the latest version here https://osf.io/vbdah/.

Q9) Imagine you have performed an independent *t*-test. There are 28 participants in each condition. In the control group, the mean is 4.1 and the standard deviation is 1.1. In the experimental group, the mean is 4.8 and the standard deviation is 1.3. Use the spreadsheet from https://osf.io/vbdah/ to determine what the effect size Hedges' *g* is.

A) 0.57

B) 0.58

C) 0.59

D) 0.66

Q10) You read a paper in the literature. All the results that are provided for an independent *t*-test are the *t*-value of 2.9, and the total sample size of 244. Use the spreadsheet to determine the best estimate of the effect size Cohen's *d* you can make based on the available data:

A) 0.004

B) 0.37

C) 0.24

D) 0.60

For further reading about effect size estimates, see this practical primer I have written.