Testing Range Predictions

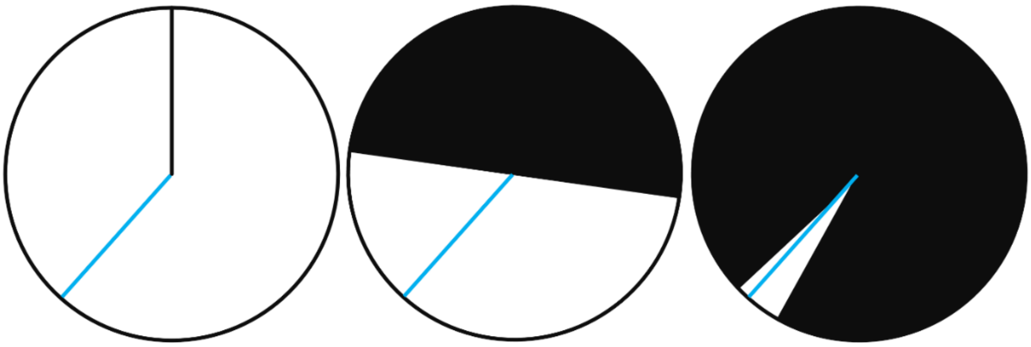
The goal of a hypothesis test is to carefully examine whether predictions that are derived from a scientific theory hold up under scrutiny. Not all predictions we can test are equally exciting. For example, if a researcher asks two groups to report their mood on a scale from 1 to 7, and then predicts the difference between these groups will fall within a range of -6 to +6, we know in advance that it must be so. No result can **falsify** the prediction, and therefore finding a result that **corroborates** the prediction is completely trivial and a waste of time.

To demonstrate our theory has good predictive validity, we need to divide all possible states of the world into a set we consider to be predicted by our theory, and a set that we do not consider predicted by our theory. We can then collect data, and if the results are in line with our prediction (repeatedly, across replication studies), our theory gains **verisimilitude** – it seems to be related to the truth. We can never know the truth, but by corroborating theoretical predictions, we can hope to get closer to it.

The most common division of states of the world that are predicted and that are not predicted by a theory in null-hypothesis significance testing is the following: An effect of exactly zero is *not* predicted by a theory, and all other effects are taken to corroborate the theoretical prediction. Here, I want to explain why this is a very weak hypothesis test. In certain lines of research, it might even be a pretty trivial prediction. Luckily, it is quite easy to perform much stronger tests of hypotheses. I'll also explain how to do so in practice.

Risky Predictions

Take a look at the three circles below. Each circle represents all possible outcomes of an empirical test of a theory. The blue line illustrates the state of the world that was observed in a (hypothetical) perfectly accurate study. The line could have fallen anywhere on the circle. We performed a study and found one specific outcome. The black area in the circle represents the states of the world that will be interpreted as *falsifying* our prediction, whereas the white area illustrates the states in the world we predicted, and that will be interpreted as *corroborating* our prediction.



In the figure on the left, only a tiny fraction of states of the world will falsify our prediction. This represents a hypothesis test where only an infinitely small portion of all possible states of the world is not in line with the prediction. A common example is a two-sided null-hypothesis significance test, which forbids (and tries to reject) only the state of the world where the true effect size is exactly zero.

In the middle circle, 50% of all possible outcomes falsify the prediction, and 50% corroborates it. A common example is a one-sided null-hypothesis test. If you predict the mean is *larger than* zero, this prediction is falsified by all states of the world where the true effect is either *equal to* zero, or *smaller than* zero. This means that half of all possible states of the world can no longer be interpreted as corroborating your prediction. The blue line, or observed state of the world in the experiment, happens to fall in the white area for the middle circle, so we can still conclude the prediction is supported. However, our prediction was already slightly riskier than in the circle on the left representing a two-sided test.

In the scenario in the right circle, almost all possible outcomes are not in line with our prediction – only 5% of the circle is white. Again, the blue line, our observed outcome, falls in this white area, and our prediction is confirmed. However, now our prediction is confirmed in a very risky test. There were many ways in which we could be wrong – but we were right regardless.

Although our prediction is confirmed in all three scenarios above, philosophers of science such as Popper and Lakatos would be most impressed after your prediction has withstood the most severe test (i.e., in the scenario illustrated by the right circle). Our prediction was most specific: 95% of possible outcomes were judged as falsifying our prediction, and only 5% of possible outcomes would be interpreted as support for our theory. Despite this high hurdle, our prediction was corroborated. Compare this to the scenario on the left – almost any outcome would have supported our theory. That our prediction was confirmed in the scenario in the left circle is hardly surprising.

Systematic Noise

The scenario in the left, where only a very small part of all possible outcomes is seen as falsifying a prediction, is very similar to how people commonly use null-hypothesis significance tests. In a null-hypothesis significance test, any statistically significant effect that is not zero is interpreted as support for a theory. Is this impressive? That depends on the possible states of the world. According to Meehl, there are many situations where null-hypothesis significance tests are performed, but the true difference is highly unlikely to be exactly zero. Meehl is especially worried about research where there is room for **systematic noise**, or the **crud factor**.

Systematic noise can only be excluded in an ideal experiment. In this ideal experiment, there is perfect random assignment to conditions, and only one single thing can cause a difference, such as in a **randomized controlled trial**. Perfection is notoriously hard to achieve in practice. In any close to perfect experiment, there can be tiny factors that, although not being the main goal of the experiment, lead to differences between the experimental and control condition. Participants in the experimental condition might read more words, answer more questions, need more time, have to think more deeply, or process more novel information. Any of these things could slightly move the true effect size away from zero – without being related to the independent variable the researchers aimed to manipulate. This is why Meehl calls it *systematic* noise, and not *random* noise: The difference is reliable, but not due to something you are **theoretically interested** in.

Many experiments are not even close to perfect and consequently have a lot of room for systematic noise. And then there are many studies where there isn't even random assignment to conditions, but where data is correlational. As an example of correlational data, think about research examining differences between women and men. If we examine differences between men and women, the subjects in our study can not be randomly assigned to a condition. In such non-experimental studies, it is possible that '**everything is correlated to everything**'. Or slightly more formally (Orben & Lakens, 2019), crud can be defined as the epistemological concept that, in correlational research, all variables are connected through multivariate causal structures which result in real non-zero correlations between all variables in any given dataset.

For example, men are on average taller than women, and as a consequence it is more common for a man to be asked to pick an object from a high shelf in a supermarket, than vice versa. If we then ask men and women 'how often do you help strangers' this average difference in height has some tiny but systematic effect on their responses. In this specific case, systematic noise moves the mean difference from zero to a slightly higher value for men – but an unknown number of other sources of systematic noise are at play, and these all interact, leading to an unknown final true population difference that is very unlikely to be exactly zero.

Q1: Null-Hypothesis Significance Tests are so common we rarely think about whether they are the right question to ask. But **when you perform a null-hypothesis test, you should justify why the null-hypothesis is an interesting hypothesis to test against**. This is not always self-evident, and sometimes the null hypothesis is simply not very interesting. Look at the last paper you wrote, or are currently writing (or if you have not written a paper, look at a paper you want to build on or apply in some way). Identify the most important hypothesis tests and ask yourself whether **the null hypothesis was justified**. Was it plausible that the null hypothesis was true? And was it an interesting value to test against?

I think there are experiments that, for all practical purposes, are controlled enough to make a point null-hypothesis a valid and realistic model to test against. However, I also think that these experiments do not encompass all the current uses of null-hypothesis testing. There are many experiments where a test against a null-hypothesis is performed, while the point null-hypothesis is not reasonable to entertain, and we can't expect the difference to be exactly zero.

In those studies (e.g., as in the experiment examining differences between men and women above) it is much more impressive to have a theory that is able to predict how big an effect is (approximately). In other words, we should aim for theories that make **point predictions**, or a bit more reasonably, given that most sciences have a hard time predicting a single exact value, **range predictions**.

Range Predictions

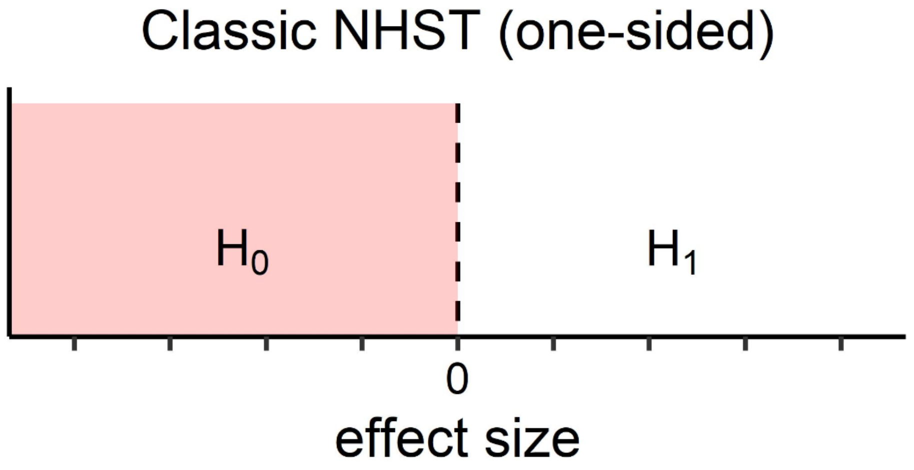
Making more risky *range predictions* has some important benefits over the widespread use of null-hypothesis tests. These benefits mean that even if a null-hypothesis test is defensible, it would be preferable if you could test a range prediction.

Making a more risky prediction gives your theory higher **verisimilitude**. You will get more credit in darts when you correctly predict you will hit the bullseye, than when you correctly predict you will hit the board. Many sports work like this, such as figure skating or gymnastics. The riskier the routine you perform, the more points you can score, since there were many ways the routine could have failed if you lacked the skill. Similarly, you get more credit for the predictive power of your theory when you correctly predict an effect will fall within 0.5 scale points of 8 on a 10 point scale, than when you predict the effect will be larger than the midpoint of the scale. A theory allows you to make predictions, and a good theory allows you to make precise predictions.

Range predictions allow you to design a study that can be **falsified based on clear criteria**. If you specify the bounds within which an effect should fall, any effect that is either smaller or larger will falsify the prediction. For a traditional null-hypothesis test, a significant effect of 0.0000001 will officially still fall in the possible states of the world that support the theory. However, it is practically impossible to falsify such tiny differences from zero, because doing so would require huge resources.

Directional Tests

Researchers often have a directional hypothesis when comparing two groups (e.g., the reaction times in the implicit association test are slower in the incongruent block compared to the congruent block). In these situations, researchers can choose to use either a two-sided test or a one-sided test. One-sided tests are more powerful than two-sided tests. If you design a test with 80% power, a one-sided test requires approximately **78% of the total sample of a two-sided test**. This means that the use of one-sided tests would make researchers more efficient. In addition, a one-sided test is a riskier prediction than a two-sided test. We have limited all possible outcomes that we predict by 50%, which is quite impressive.

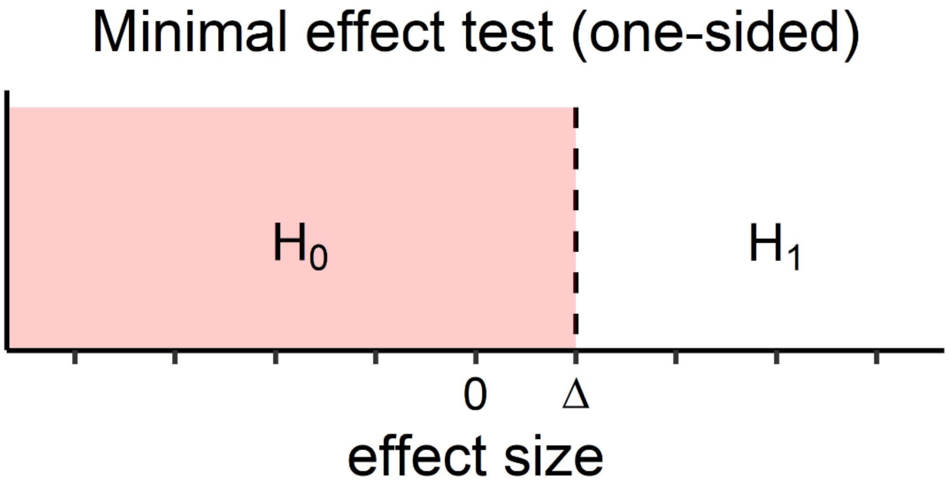


Many researchers have reacted negatively to the “widespread overuse of two-tailed testing for directional research hypotheses tests” (Cho & Abe, 2013). Others argue that directional tests should not be used. I think a fair summary of this discussion is that 1) directional tests should always be pre-registered (I agree), 2) they require smaller sample sizes, and therefore you will end up with less evidence (which is true, but if you want a specific amount of evidence, you should design studies to achieve a desired level of evidence instead of designing studies where error rates are controlled), and 3) when results in both directions are practically relevant, such as in medical research where we care both about improving lives, and not making lives worse, directional tests might only be desirable in very specific circumstances (I agree). Although these caveats are important, in many cases researchers make directional predictions. They would consider their predictions proven wrong by an effect of 0, or an effect in the opposite direction. Effects of 0, or effects in the opposite direction, might be interesting enough to follow up on. But if the question is whether some manipulation leads to a positive effect, a result of a one-sided test is the logical answer to that question.

Q2: Two-sided tests are so common we rarely think about whether they are the right question to ask. But if you make a directional prediction, it often makes sense to perform a directional test. Look at the last paper you wrote or are currently writing (or if you have not written a paper, look at a paper you want to build on or apply in some way). Identify the main hypothesis in the introduction. Did the hypothesis make a one-sided prediction? Take a look at the result section. Was the hypothesis tested in a two-sided test?

Minimal Effect Tests

A directional test can use a null-hypothesis of an effect of 0, but we can also perform a hypothesis test not against 0, but against a smallest effect size of interest. This is known as a minimal effect test. In a minimal effect test, the null-hypothesis is any value smaller than the values we care about. For example, imagine we have designed an after-school training program to improve the language ability of young children. This program has a positive effect whenever we can reject the null hypothesis of an effect size of 0. But the training program also has costs, and if it improved language ability 0.00001 on a standardize test where students can score between 0 and 100, it will not be worth implementing the training program. Based on a cost-benefit analysis, a team of experts has decided the training program is worth the costs if the improvement is larger than 5%. Therefore, they test against a smallest effect of interest (Δ) of 5, instead of testing against 0. The null hypothesis is now any effect up to 5%. We reject the null hypothesis if the observed effect is statistically larger than 5%.

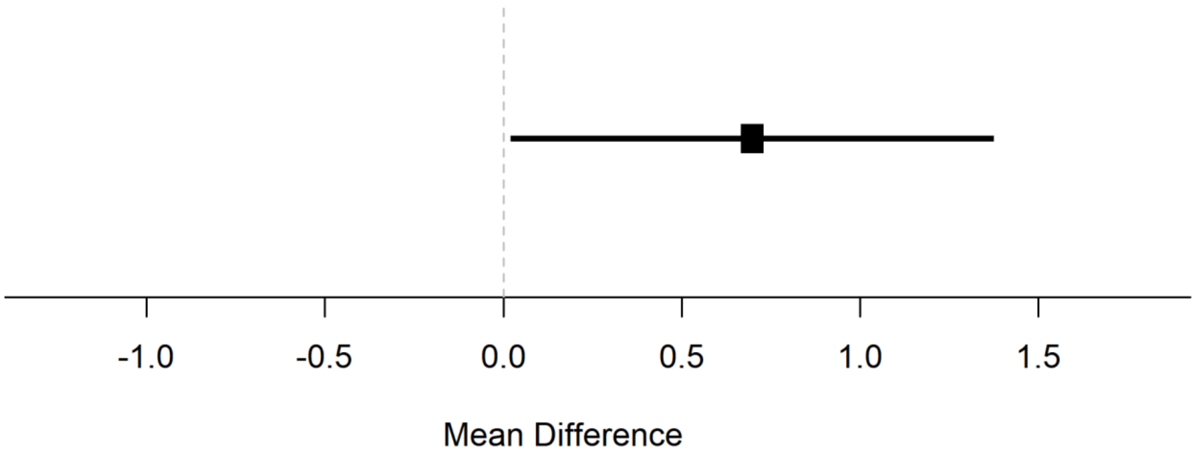


Many of the criticisms on *p*-values in null-hypothesis tests where H0 = 0 disappear when *p*-values are calculated for a minimal effect test. In a traditional hypothesis test with at least some systematic noise (meaning the true effect differs slightly from zero) all studies where the null is not exactly true will lead to a significant effect with a large enough sample size. This makes it a boring (i.e., not risky) prediction, and we will end up stating there is a 'significant' difference for tiny irrelevant effects. I expect this problem will become more important now that it is easier to get access to Big Data.

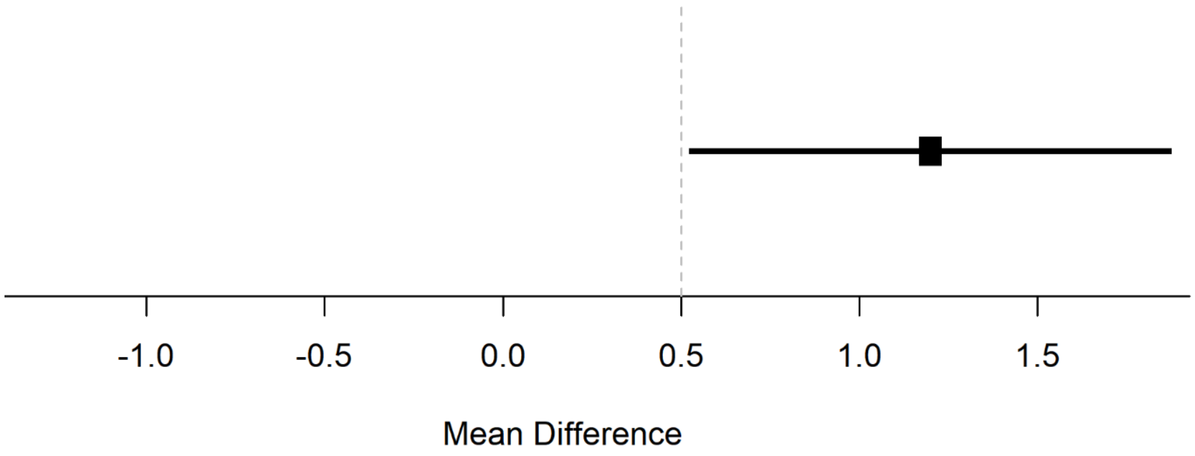
However, we don't want just any effect to become statistically significant – we want **theoretically or practically relevant** effects to be significant, but not **theoretically or practically irrelevant** effects. A minimal effect test achieves this. If we predict effects larger than 5%, an effect of 1% might be statistically different from 0 in a huge sample, but it is not **practically relevant**.

Testing Range Predictions

In a null-hypothesis test (visualized below) we compare the observed mean (the black square) and the 95% confidence interval (the length of the horizontal line through the square) against the hypothesis that the difference is 0 (indicated by the dotted vertical line at 0). Let's imagine the test yields a $p = 0.047$. If we use an alpha level of 0.05, this is just below the alpha threshold. The observed difference (indicated by the square) has a confidence interval that ranges from almost close to 0 to 1.4. We can reject the null, but beyond that, we haven't learned much.



In the example above, we were testing against a mean difference of 0. But there is no reason why a hypothesis test should be limited to test against a mean difference of 0. For example, let's assume effects smaller than 0.5 are considered too small to matter. In this case, we can perform a minimal effect test against 0.5 instead of 0. In the figure below, we again see a $p = 0.047$ result, but now for a much riskier test, namely against a smallest effect of interest of 0.5.



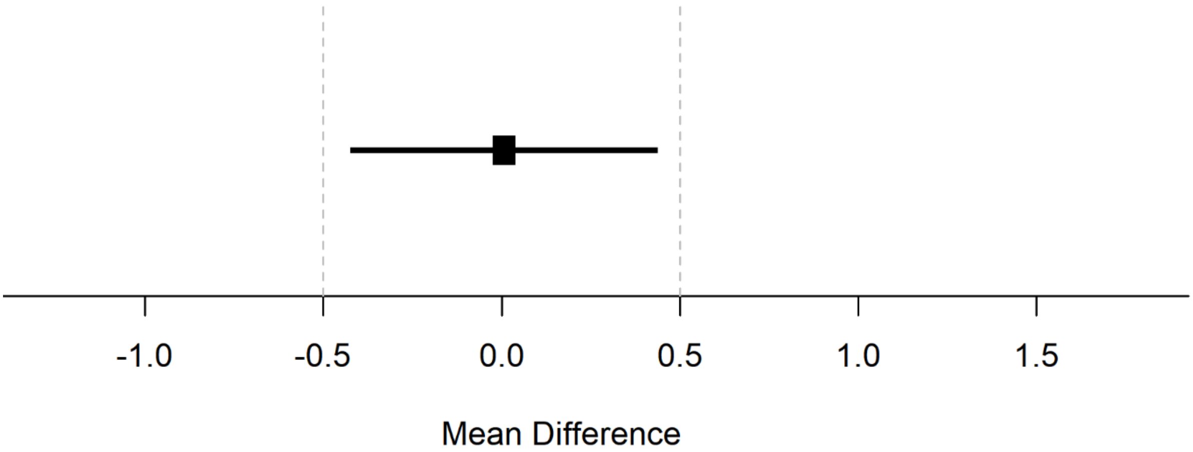
People often report a manipulation check in their articles. For example, based on previous work, they have selected 20 positive words and 20 negative words, and use these in an experiment. They might ask participants after the study to evaluate these words as a manipulation check. For example, in one of the articles that were part of my PhD thesis, I wrote:

"*Manipulation check*. Positive words were judged as more positive (*M* = 6.51) than negative words (*M* = 1.80), and a paired-samples *t*-test indicated this difference was significant, *t*(32) = 29.06, *p* < .001."

The positive and negative stimuli were evaluated on a 7-point scale. **Given that they were explicitly selected to be extremely positive and negative, is this test really contributing something**? You might argue that at least it confirms that they differ, but in practice, we are reaching a foregone conclusion. Are we happy when the difference in evaluation is simply greater than zero? Imagine I repeat the experiment with 2000 participants, and used positive words and negative words that differed statistically from each other. However, the manipulation check shows the mean of positive words is 4.10, and the mean of negative words is 3.90. I argue that the difference between words is again statistically significant. Is this a valid replication? Probably not. The real question was perhaps not if the evaluations of these two groups differ, but **how much they minimally need to differ to lead to the predicted effects**. Is a difference of 6.51-1.80 = 4.71 scale points needed? Is a difference of 0.20 scale points sufficient as well?

Q3: Look at the last paper you wrote or are currently writing (or if you have not written a paper, look at a paper you want to build on or apply in some way). Take a look at the null-hypothesis tests (whether Bayesian or frequentist) and judge if the question of whether the effect is zero is interesting, or if the authors (or you yourself) might actually have been implicitly arguing to the presence of some unspecified minimal effect. **If you were able to specify this minimal effect, would it have made more sense to report a minimal effect test for some of the hypothesis tests**?

Meehl (1967 – yes, that is more than 40 years ago!) compared the use of statistical tests in psychology and physics, and notes that in physics, researchers make point predictions. One way to test point predictions is to examine whether the observed mean falls between an upper and lower bound. Such a test is visualized in the figure below. We have set bounds of -0.5 and 0.5, and predict our observed mean falls within these bounds. Note that the bounds happen to be symmetric around 0, but you can set the bounds wherever you like.



If you have learned about **equivalence testing** (see Lakens, Scheel, & Isager, 2018, and later in this course), you might recognize the practice of specifying these bounds (referred to as equivalence bounds) to examine whether the effect falls within an equivalence range – a range of values close enough to 0 to find the effects too small to matter. An equivalence test is basically a specific version of a range prediction, where the goal is to reject effects that are large enough to matter, so that we can conclude the effect is **practically equivalent to zero**.

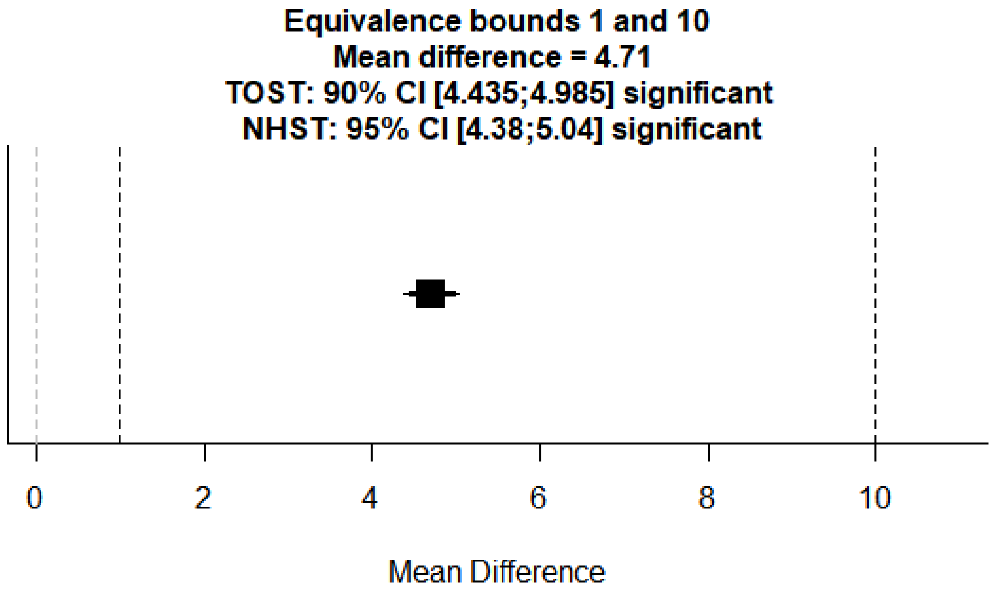
But you can use equivalence tests to test any range. Let's revisit our manipulation check example from question 3. The difference in means (in a dependent *t*-test) was 4.71 (6.51 - 1.80). We can test this difference against 0 in a one-sample t-test (which is the same as testing the two means against each other in a dependent *t*-test). But we can also perform a minimal effect test. Let's start with the modest prediction that the difference in means should be not just larger than 0, but larger than 1. In other words, we aim to test the difference score against a lower bound of 1. We have no upper bound we want to test against, and in the test below, we simply fill in a very large value (a difference of 10, while the maximum difference on a 7 point scale is 6) so that the test result is mainly determined by the one-sided test against a difference of 1 (the equivalence test used here reports the highest *p*-value from the pair of two one-sided tests).

Q4: **Open the file testing_range_predictions.R**. Run the code. You will see the following test and figure as output:

```
Equivalence Test Result:
The equivalence test was significant, t(32) = 22.892, p = 0.00000000000000
000000104, given equivalence bounds of 1.000 and 10.000 (on a raw scale) a
nd an alpha of 0.05.

Null Hypothesis Test Result:
```

```
The null hypothesis test was significant, t(32) = 29.062, p = 0.0000000000
0000000000000142, given an alpha of 0.05.
```



In the figure, we see our mean difference of 4.71, and the confidence interval around it. The difference is clearly not 0 (the confidence interval does not overlap with the vertical dashed grey line at 0), but the difference is also statistically larger than 1 (and smaller than 10, but we are not interested in the upper bound in this specific case). We see the *t*-value for the null hypothesis test (29.06) is larger than the *t*-value for the minimal effect test (22.892). Why?

A) The minimal effect test is **one-sided**, but the null hypothesis significance test is **two-sided**, and therefore the t-value for the null-hypothesis test is larger.

B) The minimal effect test is **two-sided**, but the null hypothesis significance test is **one-sided**, and therefore the t-value for the null-hypothesis test is larger.

C) The minimal effect test is **against a value of 1, which is closer to the observed mean difference,** and therefore the test result is less extreme.

D) A null-hypothesis test is uniformly most powerful – any alternative test will always have less power, and thus a lower t-value, than a null-hypothesis test.

Q5: Realistically speaking, we probably do not want the positive and negative words to just differ 1 scale point. Let's say we want the two groups of words to have evaluations that differ at least 3 scale points. Change the lower bound (or minimum effect) we want to test against in line 8 from 1 to 3. How does this affect the null-hypothesis test? How does this affect the equivalence test against the lower bound of 3?

A) The *t*-value for both the null-hypothesis test as the equivalence test are now 10.551. Adjusting the minimum effect thus influences both test results.

B) The *t*-value for the null-hypothesis test is now 10.551, the *t*-value for the equivalence test remains stable at 22.892. Adjusting the minimum effect thus influences only the null hypothesis test.

C) The *t*-value for the null-hypothesis test remains 29.06, the *t*-value for the equivalence test is lowered to 10.551. Adjusting the minimum effect thus influences only the equivalence test.

D) The *t*-value for the null-hypothesis test is now 10.551, the *t*-value for the equivalence test increases to 32.892. Adjusting the minimum effect thus influences only the equivalence test.

Q6: This is a more difficult insight question. Based on the previous two questions, a general pattern emerges. (Hint: think about effect size, and the difference we are testing in the two tests). Compared to a null hypothesis test, a minimal effect test:

A) is **riskier**, but also has **less** power, and thus a minimal effect test requires a **larger** number of observations than a null-hypothesis test.

B) is **riskier**, but also has **more** power, and thus a minimal effect test requires a **smaller** number of observations than a null-hypothesis test.

C) is **less risky**, but also has **less** power, and thus a minimal effect test requires a **larger** number of observations than a null-hypothesis test.

D) is **less risky**, but also has **more** power, and thus a minimal effect test requires a **smaller** number of observations than a null-hypothesis test.

Although Meehl prefers **point predictions that lie within a certain range**, he doesn't completely reject the use of null-hypothesis significance testing. When he asks 'Is it ever correct to use null-hypothesis significance tests?' his own answer is 'Of course it is' (Meehl, 1990). **There are times, such as very early in research lines, where researchers do not have good enough models, or reliable existing data, to make point or range predictions**. Other times, two competing theories are not more precise than that one predicts rats in a maze will learn *something*, while the other theory predicts the rats will learn *nothing*. As Meehl writes: "When I was a rat psychologist, I unabashedly employed significance testing in latent-learning experiments; looking back I see no reason to fault myself for having done so in the light of my present methodological views."

There are no good or bad statistical approaches – all statistical approaches are just answers to questions. **What matters is asking the best possible question**. It makes sense to allow traditional null-hypothesis tests early in research lines, when theories do not make more specific predictions than that 'something' will happen. But we should also push ourselves to develop theories that make more precise range predictions, and then test these more specific predictions. More mature theories should be able to predict effects in some range – even when these ranges are relatively wide.

References

Cho, H.-C., & Abe, S. (2013). Is two-tailed testing for directional research hypotheses tests legitimate? Journal of Business Research, 66(9), 1261–1266. https://doi.org/10.1016/j.jbusres.2012.02.023

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science*, 2515245918770963. https://doi.org/10/gdj7s9

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. Philosophy of Science, 103–115.

Meehl, P. E. (1990). Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it. *Psychological Inquiry*, *1*(2), 108–141.

Orben, A., & Lakens, D. (2019). Crud (Re)defined. https://doi.org/10.31234/osf.io/96dpy