Brandt et al. example: equivalent & not significant

Equivalence bounds based on critical effect sizes

Another justifiable choice we would like to propose is to use the smallest observed effect size that could have been statistically significant in the original study. Based only on the alpha level and the sample size, we can calculate the criticial test value (e.g., t, F, Z). This critical test value can be transformed to a standardized effect size (e.g., $d_{crit} = t_{crit}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$), which can thus be interpreted as a *critical effect size*[1]. Observed effect sizes smaller than the critical effect size would not have been statistically significant in the original study. Based on this we can guess that the authors were not interested in effects smaller than this critical effect size, and thus use it as the SESOI. An equivalence test with these bounds can reject all observed effect sizes that the original study had the power to detect.

As an example, Banerjee, Chatterjee, & Sinha (2012) reported that participants who had been asked to describe an unethical deed from their past judged the room to be darker than participants who had been asked to describe an ethical deed ($M_{unethical} = 4.71$, $SD_{unethical} = 0.85$, $M_{ethical} = 5.3$, $SD_{ethical} = 0.97$, $t(38) = 2.03$, $p = 0.049$, $d = 0.65$). A close replication by Brandt, IJzerman, & Blanken (2014) found no significant effect ($t(98) = 0.56$, $p = 0.574$, $d = 0.11$). The smallest effect the original study could have detected is $d_{crit} = 2.02\sqrt{\frac{1}{20} + \frac{1}{20}} = 0.64$. Using this as our SESOI for a TOST with Welch's t-test for independent samples — resulting in equivalence bounds of $\Delta_L = -0.64$ and $\Delta_U = 0.64$ — we indeed find that the effect reported by the replication study is equivalent, $t(97.78) = -2.64$, $p = 0.005$.

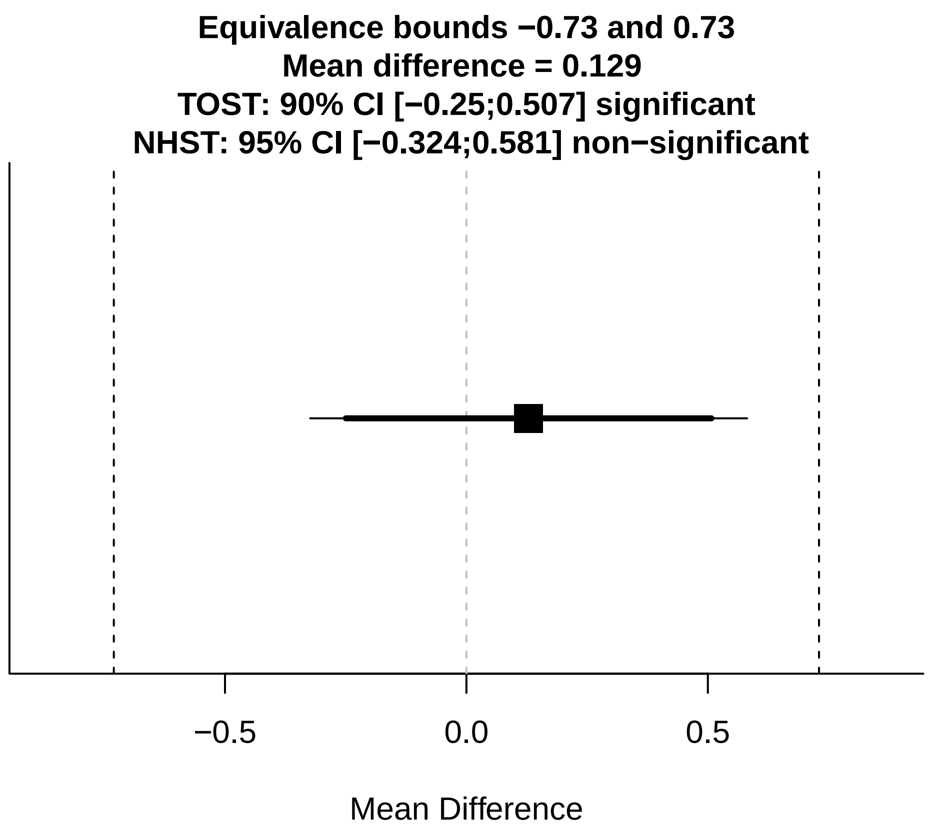
```
## Using alpha = 0.05 Welch's t-test was non-significant, t(97.78283) = 0.5649507, p = 0.5734012

##

##
## Using alpha = 0.05 the equivalence test based on Welch's t-test  was significant, t(97.78283) = -2.63

##
```

[1]This will typically, although not always, correspond to the effect size the study had 50% power to detect. This procedure will thus result in effect sizes that are substantially larger than the ones obtained using the small telescopes approach, which gives the effect size a study had 33% power to detect.




```
## TOST results:
##   t-value 1    p-value 1 t-value 2   p-value 2       df
## 1   3.76796 0.0001407233 -2.638058 0.004851125 97.78283
##
## Equivalence bounds (Cohen's d):
##   low bound d high bound d
## 1  -0.6401696    0.6401696
##
## Equivalence bounds (raw scores):
##   low bound raw high bound raw
## 1    -0.7302364      0.7302364
##
## TOST confidence interval:
##   Lower Limit 90% CI raw Upper Limit 90% CI raw
## 1              -0.249788               0.507388
```