Equivalence Tests in ARM

Jun 9, 2023

Overview

- Quoting from [1] D. G. Altman and J. M. Bland. Absence of evidence is not evidence of absence. BMJ, 311:485, 1995:

- By convention, a P value greater than 5% (P>0.05) is called “not significant”. Randomized controlled clinical trials that do not show a significant difference between the treatments being compared are often called “negative”. This term wrongly implies that the study has show that there is no difference, whereas usually all that has been shown is an absence of evidence of a difference. These are quite different statements.

Overview

- Although the quote in the previous slide addresses randomized controlled trials specifically, the same can be said for trials analyzed using ARM.

- ARM provides mean separation letters, with the following caption
  - `Means followed by the same letter or symbol do not significantly differ`
- usually followed by a note of the critical P value and mean comparison test. ARM further allows a special symbol to be displayed when no significant difference between treatment means is detected.

Null Hypothesis Significance Tests

- When we talk about significance in this context, we are usually talking about a the statistical significance of the null hypothesis.
- Remember, the null hypothesis postulates that treatment means are equal, i.e.
    - $H_0 : \mu_i = \mu_j$
- When we fail to reject the null hypothesis, we frequently state the means are not significantly different.
- But we'll not have proven the null hypothesis to be true. It's asserted to be true, then we test how well the data conforms to the null hypothesis.

NHST

- When we start an experiment, we generally don't believe the null hypothesis to be true - if it were, we wouldn't be doing the experiment. In general, we design experiments to determine if treatments are different.

- But suppose we want to design an experiment to show that two treatments are equivalent.

    - We might, for example, have a new formulation of a standard agronomic treatment. We'll refer to the mean of the new formulation to be μ_{new}, and the older standard treatment to be μ_{std}

Two one-sided tests (TOST)

- Now, it would be difficult to prove that two means are exactly equal, considering that we usually work with mean estimates that represent continuous random variables.

- Instead, we define an equivalent limit or bound (Δ), such that the difference between means is less than the limit ($|\mu_{new} - \mu_{std}| < \Delta$). This leads us to two one-tailed hypothesis (following Shuirmann, 1987)
  - $H_1\colon \mu_{new} - \mu_{std} \leq \Delta_L$
  - $H_2\colon \mu_{new} - \mu_{std} \geq \Delta_U$

- D. J. Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. Journal of Pharmacokinetics and Biopharmaceutics, 15(6), 1987.

TOST

- If we reject both hypothesis,
  - $H_1: \mu_{new} - \mu_{std} \leq \Delta_L$
  - $H_2: \mu_{new} - \mu_{std} \geq \Delta_U$
- then we can assert that $\Delta_L < \mu_{new} - \mu_{std} < \Delta_U$; or that the two treatment means are equivalent.

TOST t-tests

- The dual hypothesis can be rewritten as
    - $H_1 : (\mu_{new} - \mu_{std}) - \Delta_L \leq 0$
    - $H_2 : (\mu_{new} - \mu_{std}) - \Delta_U \geq 0$

- which leads to the t statistics
    - $t_1 = \frac{(\overline{X}_{new} - \overline{X}_{std}) - \Delta_L}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ and $t_2 = \frac{(\overline{X}_{new} - \overline{X}_{std}) - \Delta_U}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$

- where $\overline{X}_{new}$ and $\overline{X}_{std}$ are the estimated means of the new and standard treatments, respectively, while n_1 and n_2 are the number of observations for new and standard treatments.

TOST t-tests

- In some examples in the literature, the t statistics are written as

    - $t_1 = \frac{(\overline{X}_{new} - \overline{X}_{std}) - \Delta_L}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ and $t_2 = \frac{\Delta_U - (\overline{X}_{new} - \overline{X}_{std})}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$

- This usually results in a change of sign of t_2, and a change of the "tail" of the calculation of probabilities of the associated t values. We use the previous convention to be consistent with the R library TOSTER to provide an alternative check on our calculations.

- [1] A. R. Caldwell. Exploring equivalence testing with the updated toster r package. PsyArXiv, November 2022.

TOST t-tests

- s is a pooled standard deviation. In the literature, this is commonly given as
  - $$s = \sqrt{\frac{(n_1-1)sd_{new}^2+(n_2-1)sd_{std}^2}{n_1+n_2-2}}$$
- where sd_{new}^2 and sd_{std}^2 are the squared standard deviations for the new and standard treatments, respectively.
- However, when computing equivalence tests in ARM for designed experiments, we use the pooled standard deviation from the AOV table; this is the square root of the residual mean square. That is,
  - $$s = \sqrt{RMS}$$

Confidence Interval Method

- The two one sided tests imply an equivalent confidence interval test.
- Using the Confidence interval method, we define a (1-2α) confidence interval; if this interval is contained within the confidence bounds we reject the TOST hypothesis and declare that equivalence is established.
- This corresponds to a 90% CI when each TOST test is at the traditional 95% significance.

- H. van der Voet, J. N. Perry, B. Amzal, and C. Paoletti. A statistical assessment of differences and equivalences between genetically modified and reference plant varieties. BMC Biotechnology, 11(1):15, Feb. 2011.

Selecting Δ

- Choosing an appropriate Δ is not a statistical problem.

- Instead, the choice of Δ is chosen to be the smallest effect size of interest (SESOI). This will usually be chosen based on what researchers consider to be an inconsequential difference.

- Note that we do not use the term “equal” to describe two means, we use the less precise term “equivalent”, which implies a degree of uncertainty in the comparison.

Specifying Δ

- In ARM, we allow three options for specifying Δ.

- The most generally useful, for most cases, will be the “Percent of standard” basis. Suppose we enter 5 (for 5% of standard) in the ARM Limit field, and further suppose that the mean for the standard treatment ($\overline{X}_{std}$) is 50. Then
  - $\Delta_L = -\overline{X}_{std} \times (5/100) = -2.5$
  - $\Delta_U = \overline{X}_{std} \times (5/100) = 2.5$

Specifying Δ

- We also allow Cohen's d to be used to specify the equivalence limit.
- Cohen's d is a measure of effect size, or the proportion to an effect to standard deviation, such that
    - $d = \left(\overline{X}_{new} - \overline{X}_{std}\right)/s$
- where s is a pooled standard deviation.
- Cohen (whose work was in social sciences) gave a guidelines for effect sizes of small = 0.10, medium = 0.3 and large = 0.5
- When using Cohen's d as a equivalence limit basis, we compute
    - $\Delta = d \times s$

Specifying Δ

- Finally, in ARM we allow an absolute value to be entered as the confidence limit. When the absolute value basis is selected and 5 entered in the Equivalence limit field. then

  - $\Delta_L = -5 and \Delta_U = 5$

- This may be easier for researchers to work with. For example, suppose a new formulation is $15 per acre cheaper when applied to soybeans. Suppose soybeans market at $8 per bushel.

- Then at absolute value of 2 bushels per acre, the new formulation will be equivalent to the standard formulation even if mean yield associated with the new formulation is no less than 2 bushel per acre lower than the standard formula ($\Delta_L = -2$)

Published Examples

- In most cases, the literature on equivalence testing is based on two independent sample t-tests.

- These examples typically use pooled standard deviations, and may use different versions of the one-sided t statistic. Where possible, we enter the data in ARM as 2 treatment CRD trials, and use the pooled standard deviation from the AOV table. Some cases are paired observations and use a paired t-test. We enter these as RCB designs, where pair is the blocking variable.

- In some cases, the SESOI is on a absolute scale. This is a reasonable approach if only one assessment is to be tested for equivalence.

- When multiple assessments are to be tested for equivalence, there are two caveats that must be consider:

    - Absolute value SESOI may be invalid for some assessment columns

    - We have no suitable method for the multiple comparison problem across assessments, so we can't control Type I error rates for multiple equivalence tests. That is, two treatments may appear to be equivalent for a large number of assessments, but some of those equivalences may be spurious.

Possible Outcomes

- There are different ways of stating the possible outcome of the null hypothesis test (NHST) for difference between means, and the TOST for equivalence of two means.

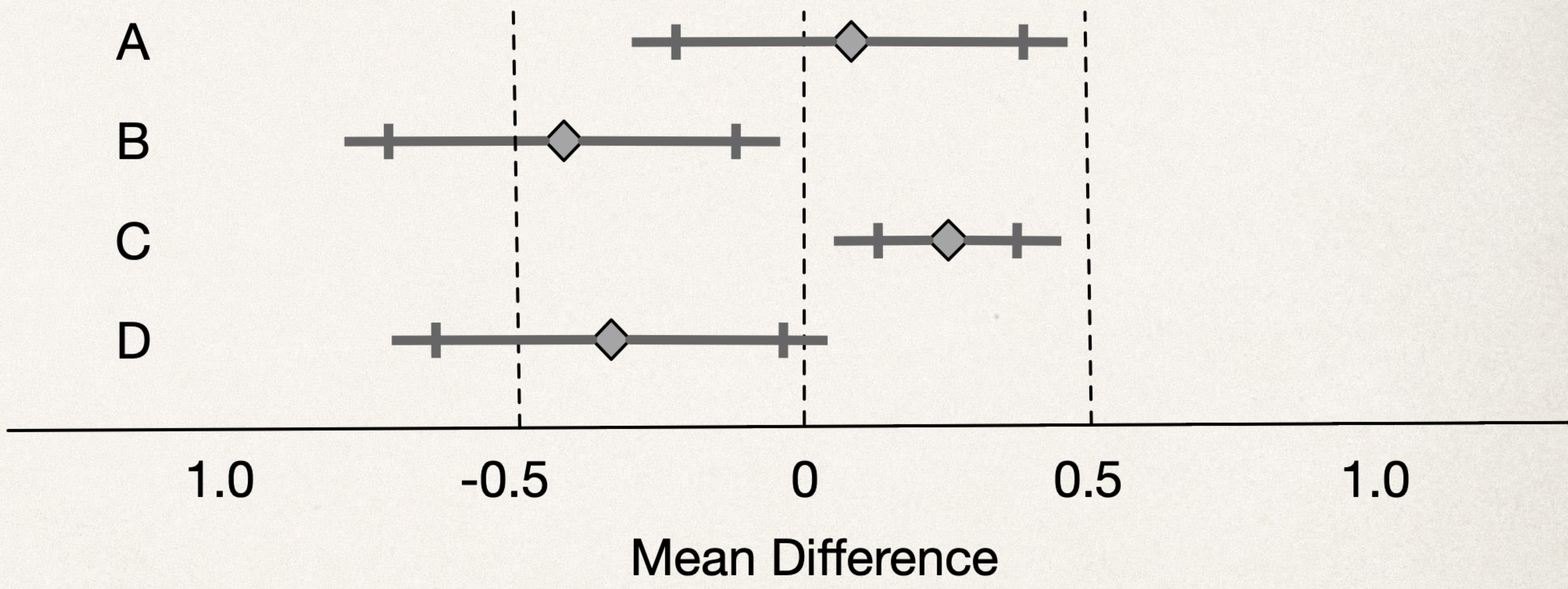
- We will borrow from Lakens, who uses confidence intervals to illustrate the outcomes.

- Four possible outcomes are presented. Diamonds represent difference between two means, while bars represent 95% confidence intervals; horizontal bands on the bars represent 90% C.I.

- Note that it is possible for two treatments to be statistically significant and statistically equivalent.(Case C). This may happen if the equivalence limit is large, and the 90% CI of the mean difference is small

- Adapted from D. Lakens. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. Social Psychological and Personality Science, 8(4):355–362, May 2017.

A	Statistically Equivalent and Not Different
B	Not Equivalent and Statistically Different
C	Statistically Equivalent and Statistically Different
D	Not Equivalent and Not Different



Lakens 2018

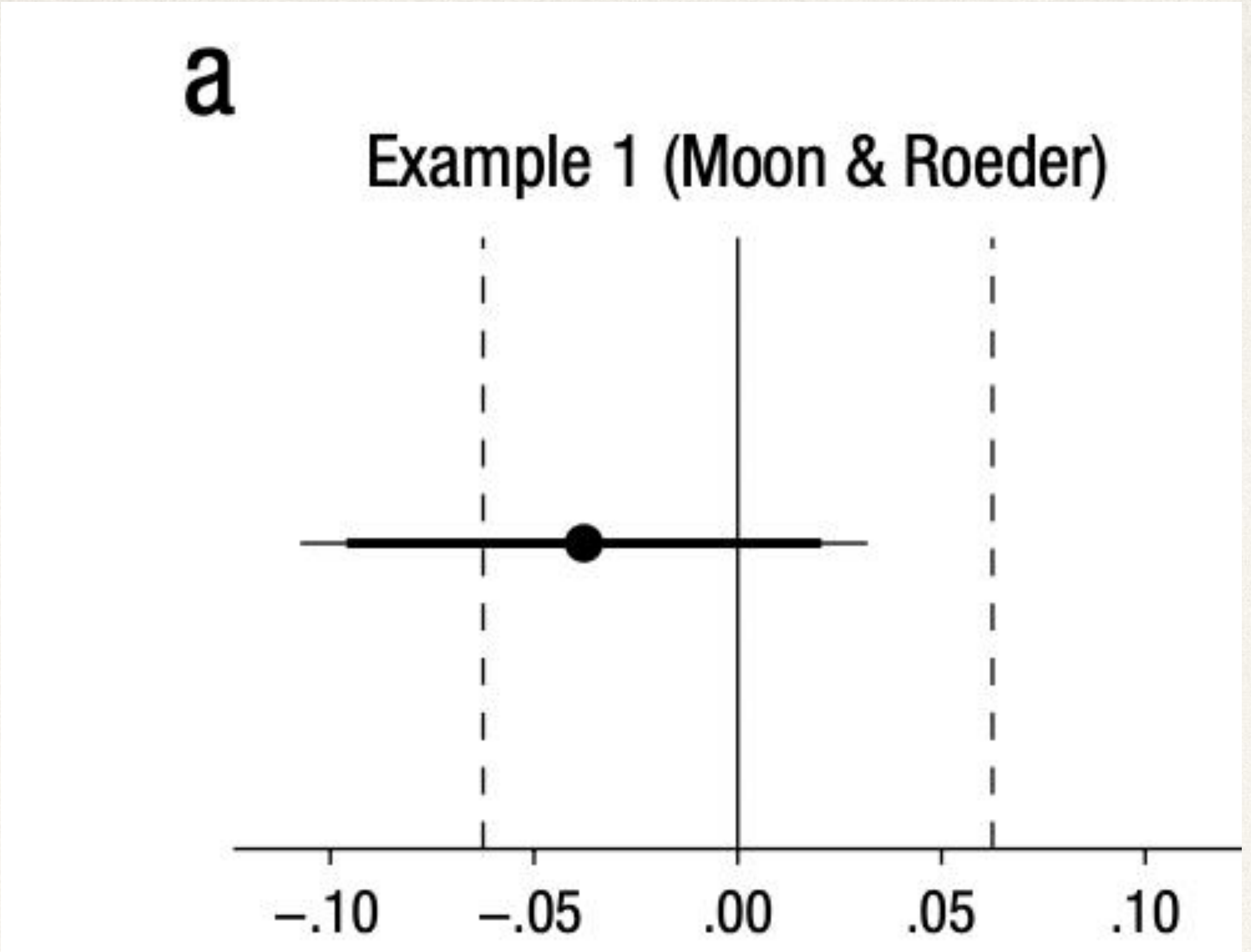
- Lakens provides several examples of cases when both NHST and TOST may be either significant or non-significant. However, they only supply data for Example 1.

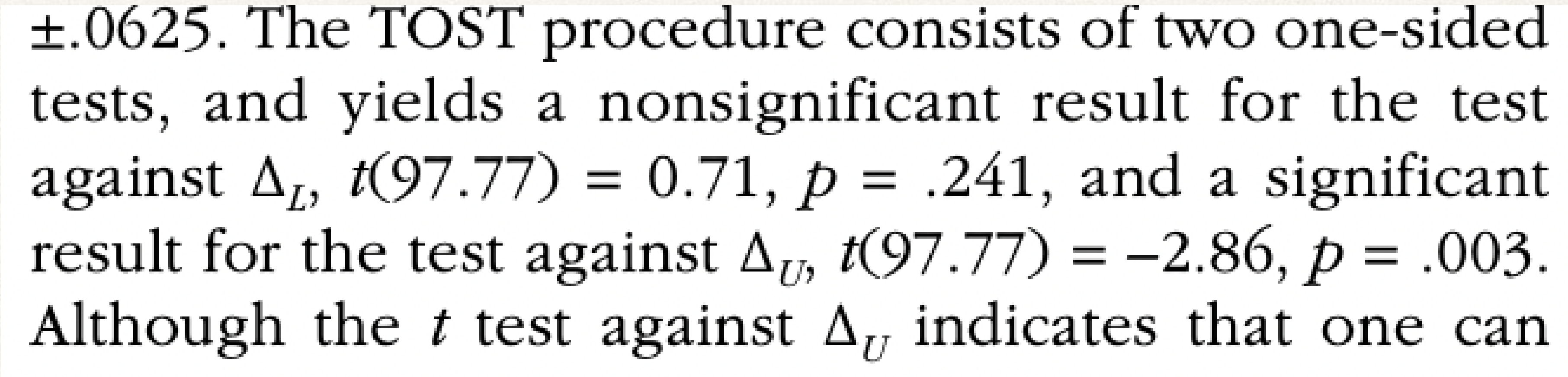
- This is an example of outcome 1 - not statistically different and not equivalent.

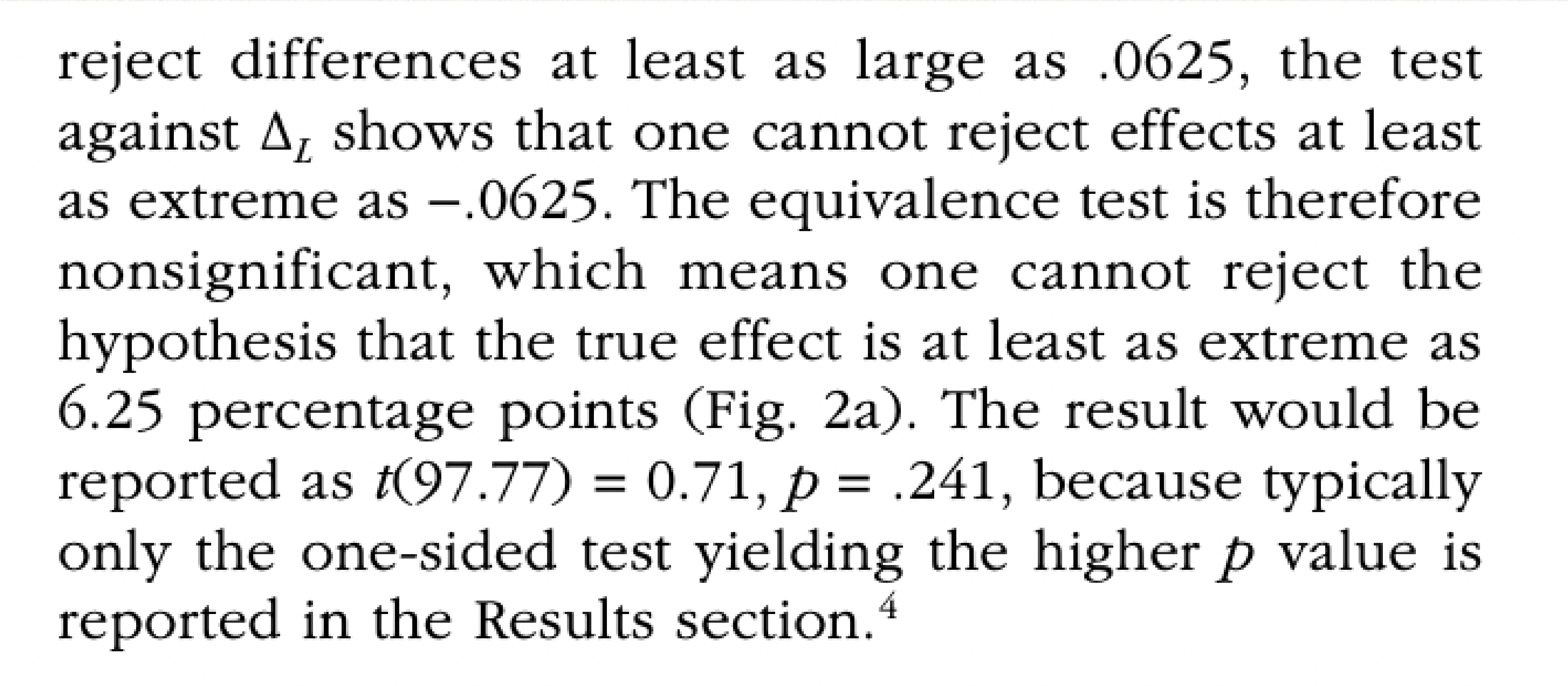
- We enter the data as a CRD experiment in ARM.

- Lakens was the original author of the TOSTER library, but additional modifications where made by Caldwell

- [1] D. Lakens, A. M. Scheel, and P. M. Isager. Equivalence testing for psychological research: A tutorial. Advances in Methods and Practices in Psychological Science, 1(2):259–269, June 2018.
- [2] A. R. Caldwell. Exploring equivalence testing with the updated toster r package. PsyArXiv, November 2022.

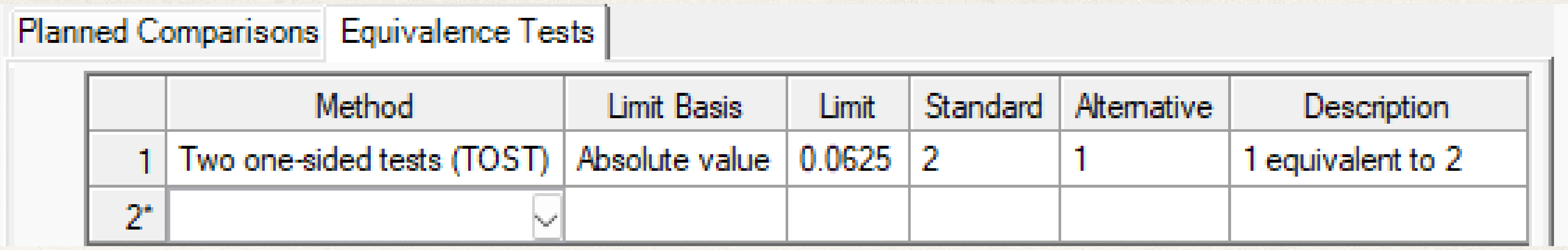



±.0625. The TOST procedure consists of two one-sided tests, and yields a nonsignificant result for the test against Δ_L, $t(97.77) = 0.71$, $p = .241$, and a significant result for the test against Δ_U, $t(97.77) = –2.86$, $p = .003$. Although the t test against Δ_U indicates that one can


reject differences at least as large as .0625, the test against Δ_L shows that one cannot reject effects at least as extreme as –.0625. The equivalence test is therefore nonsignificant, which means one cannot reject the hypothesis that the true effect is at least as extreme as 6.25 percentage points (Fig. 2a). The result would be reported as $t(97.77) = 0.71$, $p = .241$, because typically only the one-sided test yielding the higher p value is reported in the Results section.[4]

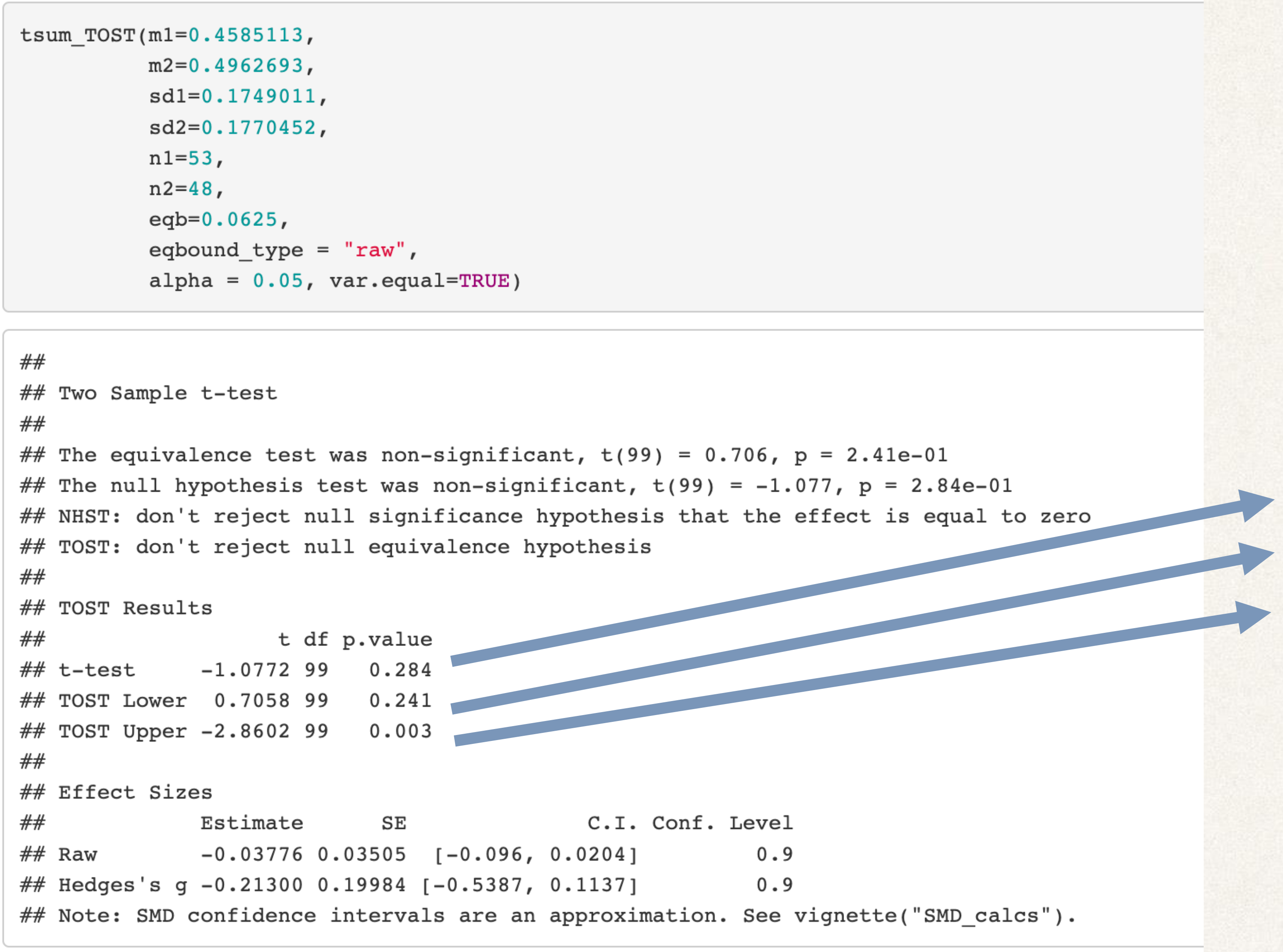
Lakens 2018

- Lakens, et. al, argue that an absolute difference between the scores of 6.25% is not meaningful.
- In ARM, we enter this as using the Absolute value as the Limit Basis, and enter 0.0625 in the Limit field. This examples used the TOST method, so we enter that as well. We also enter treatment 1 as the alternative, and 2 as the standard. Remember, we compute mean difference as $\overline{X}_{new} - \overline{X}_{std}$.
- For comparison, we enter the means and standard deviations for these data in R and invoke the tsum_TOST TOSTER library function.



	Method	Limit Basis	Limit	Standard	Alternative	Description
1	Two one-sided tests (TOST)	Absolute value	0.0625	2	1	1 equivalent to 2
2*						

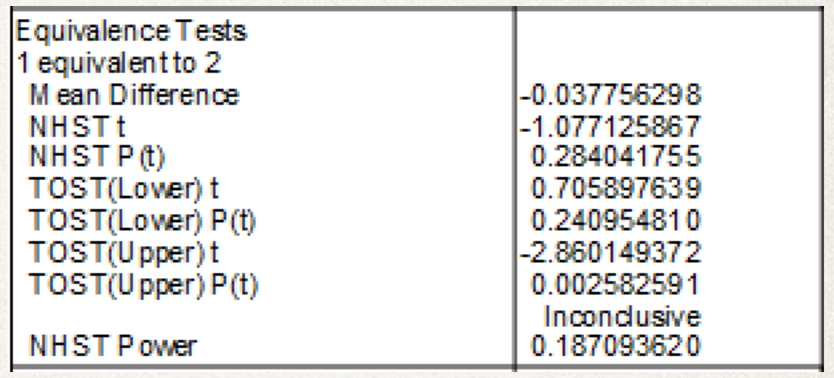
TOSTER output



```
tsum_TOST(m1=0.4585113,
          m2=0.4962693,
          sd1=0.1749011,
          sd2=0.1770452,
          n1=53,
          n2=48,
          eqb=0.0625,
          eqbound_type = "raw",
          alpha = 0.05, var.equal=TRUE)
```

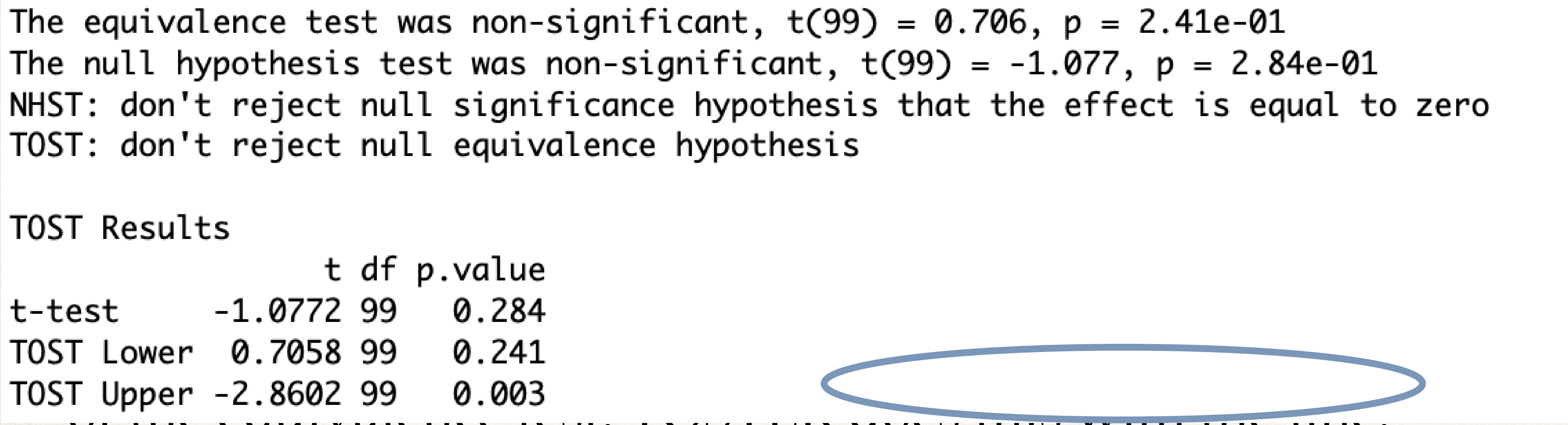
```
##
## Two Sample t-test
##
## The equivalence test was non-significant, t(99) = 0.706, p = 2.41e-01
## The null hypothesis test was non-significant, t(99) = -1.077, p = 2.84e-01
## NHST: don't reject null significance hypothesis that the effect is equal to zero
## TOST: don't reject null equivalence hypothesis
##
## TOST Results
##                   t df p.value
## t-test     -1.0772 99   0.284
## TOST Lower  0.7058 99   0.241
## TOST Upper -2.8602 99   0.003
##
## Effect Sizes
##              Estimate      SE              C.I. Conf. Level
## Raw         -0.03776 0.03505  [-0.096, 0.0204]         0.9
## Hedges's g -0.21300 0.19984 [-0.5387, 0.1137]         0.9
## Note: SMD confidence intervals are an approximation. See vignette("SMD_calcs").
```

ARM output



Equivalence Tests	
1 equivalent to 2	
Mean Difference	-0.037756298
NHST t	-1.077125867
NHST P(t)	0.284041755
TOST(Lower) t	0.705897639
TOST(Lower) P(t)	0.240954810
TOST(Upper) t	-2.860149372
TOST(Upper) P(t)	0.002582591
	Inconclusive
NHST Power	0.187093620

Lakens 2018



```
The equivalence test was non-significant, t(99) = 0.706, p = 2.41e-01
The null hypothesis test was non-significant, t(99) = -1.077, p = 2.84e-01
NHST: don't reject null significance hypothesis that the effect is equal to zero
TOST: don't reject null equivalence hypothesis

TOST Results
                 t df p.value
t-test     -1.0772 99   0.284
TOST Lower  0.7058 99   0.241
TOST Upper -2.8602 99   0.003
```

the omnibus test of equivalence.

ST with the largest p-value as result

of the equivalence test. TOSTER does this with the line:

- We require both tests to be significant to reject the pair of null hypothesis and establish equivalence.

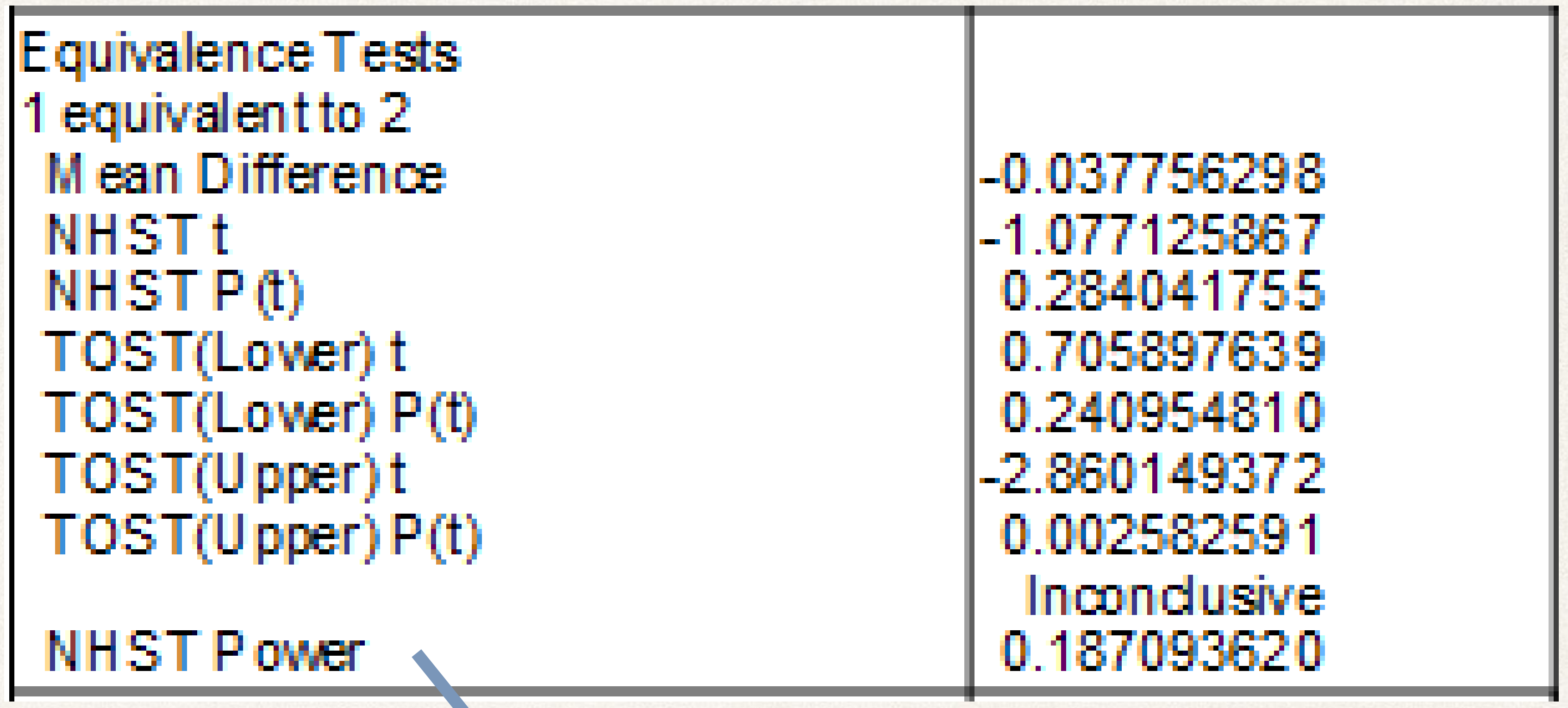
Lakens 2018

- In ARM, we use the language “Inconclusive” when we reject only one of the two null hypothesis.
- In this case, we reject the Upper TOST. Remember, this takes the form
  - H_2: $(\mu_{new} - \mu_{std}) - \Delta_U \geq 0$ or H_2: $(\mu_{new} - \mu_{std}) \geq \Delta_U$
- When we reject this hypothesis, we tend to accept the alternative, that H_{2a}: $(\mu_{new} - \mu_{std}) < \Delta_U$, or that the difference between the two means is smaller than our upper bound of equivalence.
- This may lead us to conclude that the new treatment is not superior to the standard, within the confidence limits.
- Note that we are not testing the hypothesis H_2: $(\mu_{new} - \mu_{std}) \geq 0$, which is a one-tail version of the standard NHST, which is usually two-tailed. Thus, equivalence testing is not quite the same as testing a one-tailed null hypothesis of the traditional form - the equivalence limit plays a role in inference.

Lakens 2018

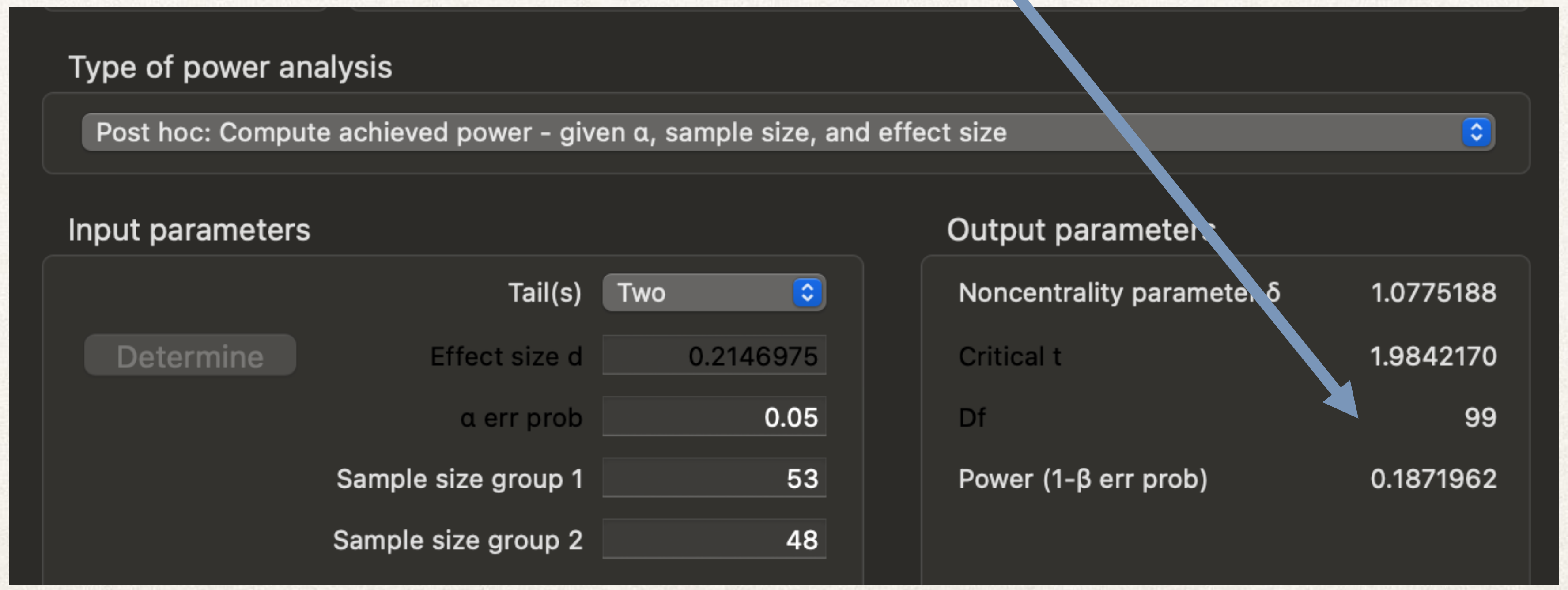
- The TOSTER library does not perform power calculations for the achieved power of the NHST, but we include this calculation in ARM.
- This tells use if we have sufficient statistical power to detect a mean different of the same magnitude as the calculated mean difference.
- We can compare with value with the Post hoc: Compute achieved power … option from the G*Power software.

ARM output



```
Equivalence Tests
1 equivalent to 2
  Mean Difference          -0.037756298
  NHST t                   -1.077125867
  NHST P(t)                 0.284041755
  TOST(Lower) t             0.705897639
  TOST(Lower) P(t)          0.240954810
  TOST(Upper) t            -2.860149372
  TOST(Upper) P(t)          0.002582591
                             Inconclusive
  NHST Power                0.187093620
```

G*Power output



Type of power analysis: Post hoc: Compute achieved power - given α, sample size, and effect size

Input parameters

- Tail(s): Two
- Effect size d: 0.2146975
- α err prob: 0.05
- Sample size group 1: 53
- Sample size group 2: 48

Output parameters

- Noncentrality parameter δ: 1.0775188
- Critical t: 1.9842170
- Df: 99
- Power (1-β err prob): 0.1871962

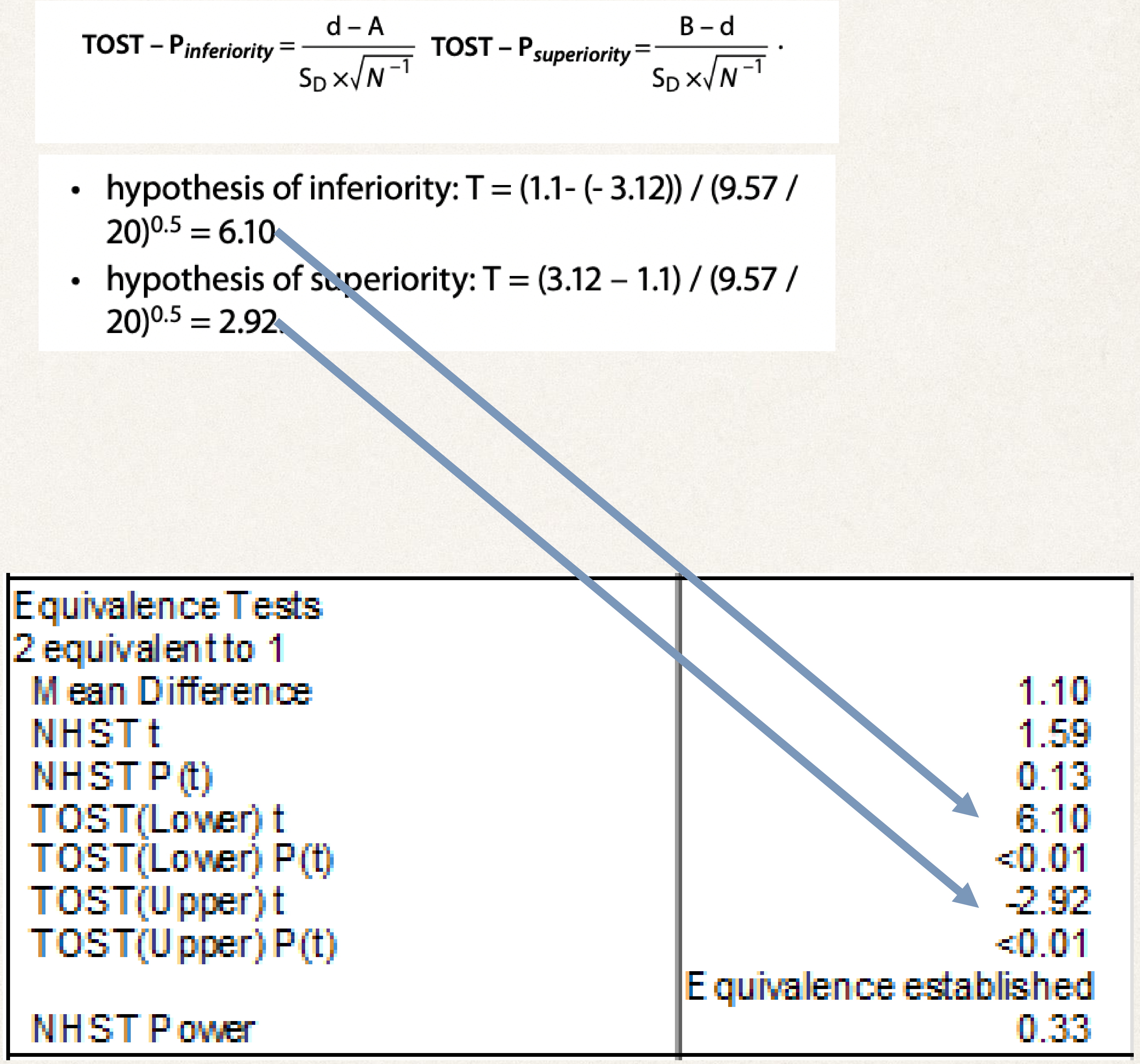
Lakens 2018

- As in this case, most ARM trials will have too few reps to have a meaningful achieved power when treatment differences are not significant. It is important to consider that in many cases true but small differences cannot be detected.

- Part of the art of selecting equivalence limits is deciding upon the value of a true difference that is functionally meaningless. We should try to select an equivalence limit to represent a difference that even if statistically significant, it would be practically insignificant.

Iolango 2017

- Ialongo (2017) provides an overview of the logic of equivalence testing, building on the logic of traditional null hypothesis testing.
- In Appendix A, Ialongo provided numbers for a paired data example. There is no experimental information associated with the values; these values are presented as examples for calculations only.
- Ialongo uses a different version of the t statistic for the upper TOST test; reversing the order of the subtractions in the numerator. This changes the sign of the t statistic, but does not change the magnitude. This does require the use of the opposite tail of the t distribution to calculate p values.
- C. Ialongo. The logic of equivalence testing and its use in laboratory medicine. Biochemia Medica, 27(1):5–13, 2017.



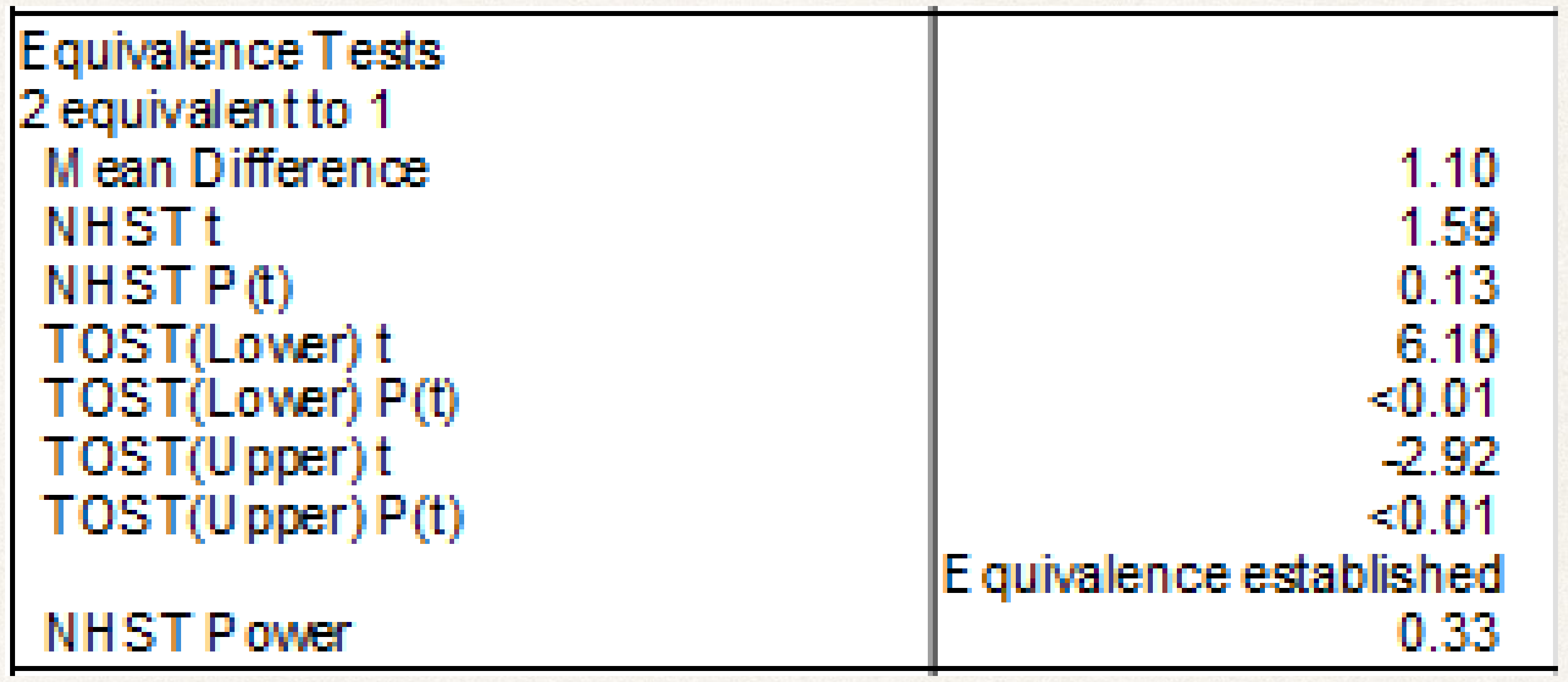
$$\text{TOST} - \text{P}_{\textit{inferiority}} = \frac{\text{d} - \text{A}}{\text{S}_\text{D} \times \sqrt{N^{-1}}} \quad \text{TOST} - \text{P}_{\textit{superiority}} = \frac{\text{B} - \text{d}}{\text{S}_\text{D} \times \sqrt{N^{-1}}}\,.$$

- hypothesis of inferiority: $T = (1.1- (- 3.12)) / (9.57 / 20)^{0.5} = 6.10$
- hypothesis of superiority: $T = (3.12 – 1.1) / (9.57 / 20)^{0.5} = 2.92$.

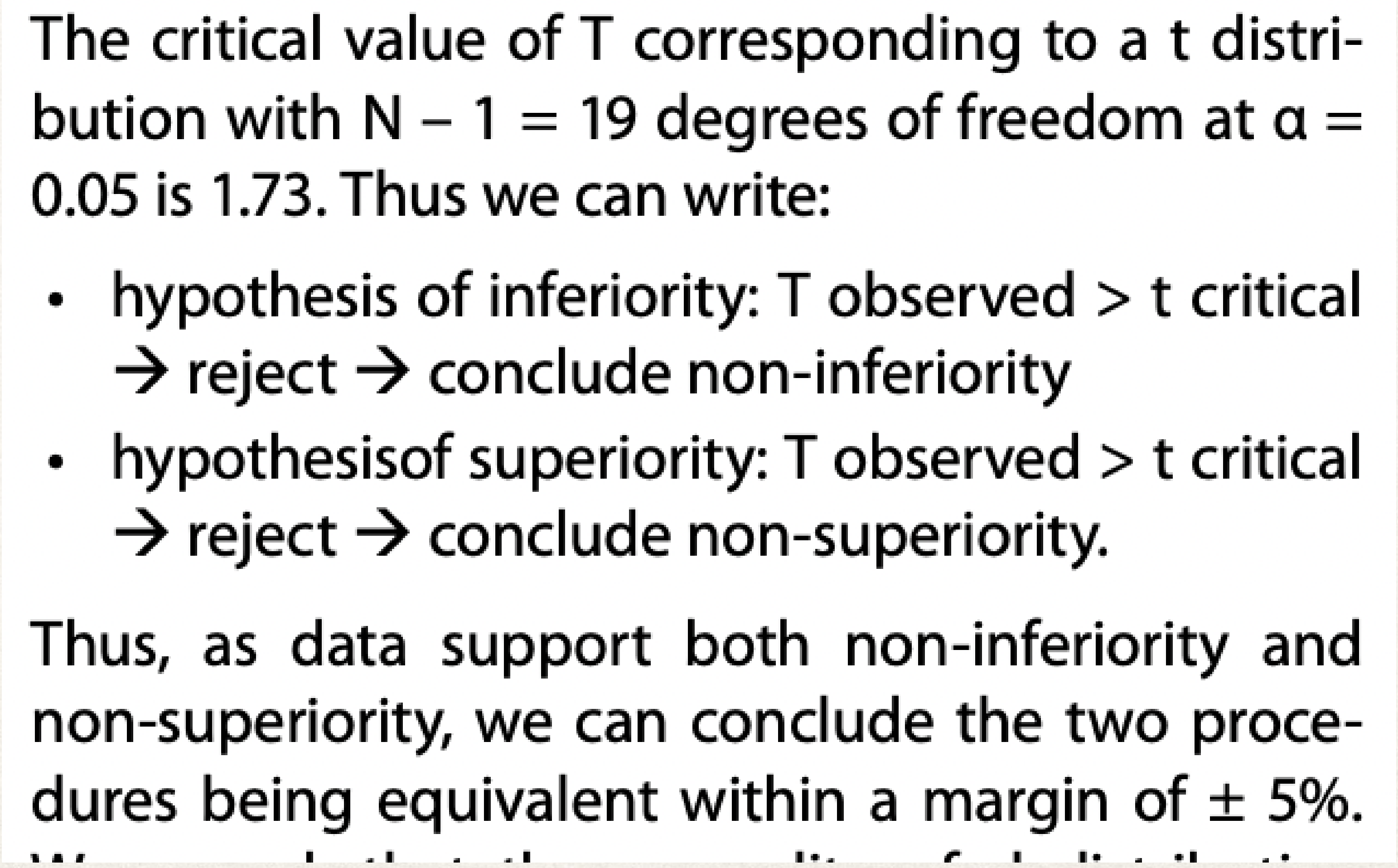
Equivalence Tests	
2 equivalent to 1	
Mean Difference	1.10
NHST t	1.59
NHST P(t)	0.13
TOST(Lower) t	6.10
TOST(Lower) P(t)	<0.01
TOST(Upper) t	-2.92
TOST(Upper) P(t)	<0.01
	Equivalence established
NHST Power	0.33

Iolango 2017

- We reject both the hypothesis of inferiority (TOST Lower) and the hypothesis of superiority (TOST Upper), thus we can conclude that the two treatments are equivalent within a margin of 5% of standard.
- Note that Ialongo uses the terminology such that TOST(Lower) = Hypothesis of inferiority and TOST(Upper) = Hypothesis of superiority.



Equivalence Tests	
2 equivalent to 1	
Mean Difference	1.10
NHST t	1.59
NHST P(t)	0.13
TOST(Lower) t	6.10
TOST(Lower) P(t)	<0.01
TOST(Upper) t	-2.92
TOST(Upper) P(t)	<0.01
	Equivalence established
NHST Power	0.33



The critical value of T corresponding to a t distribution with N – 1 = 19 degrees of freedom at α = 0.05 is 1.73. Thus we can write:

- hypothesis of inferiority: T observed > t critical → reject → conclude non-inferiority
- hypothesisof superiority: T observed > t critical → reject → conclude non-superiority.

Thus, as data support both non-inferiority and non-superiority, we can conclude the two procedures being equivalent within a margin of ± 5%.

Iolango 2017

- Ialongo also provides a graph showing the confidence interval method. We duplicate this in ARM by selecting Confidence Interval as the equivalence method.
- Since the 90% CI of treatment difference (0.18,2.02) is contained in the Equivalence interval (-3.12, 3.12) we conclude the treatments are equivalent.
- We should note that ARM output disagrees with Iolango. Iolongo's CI corresponds to critical t values at (1-2α), which produces a 80% two-tailed CI. ARM uses (1-α), which produces a 90% two-tailed CI.

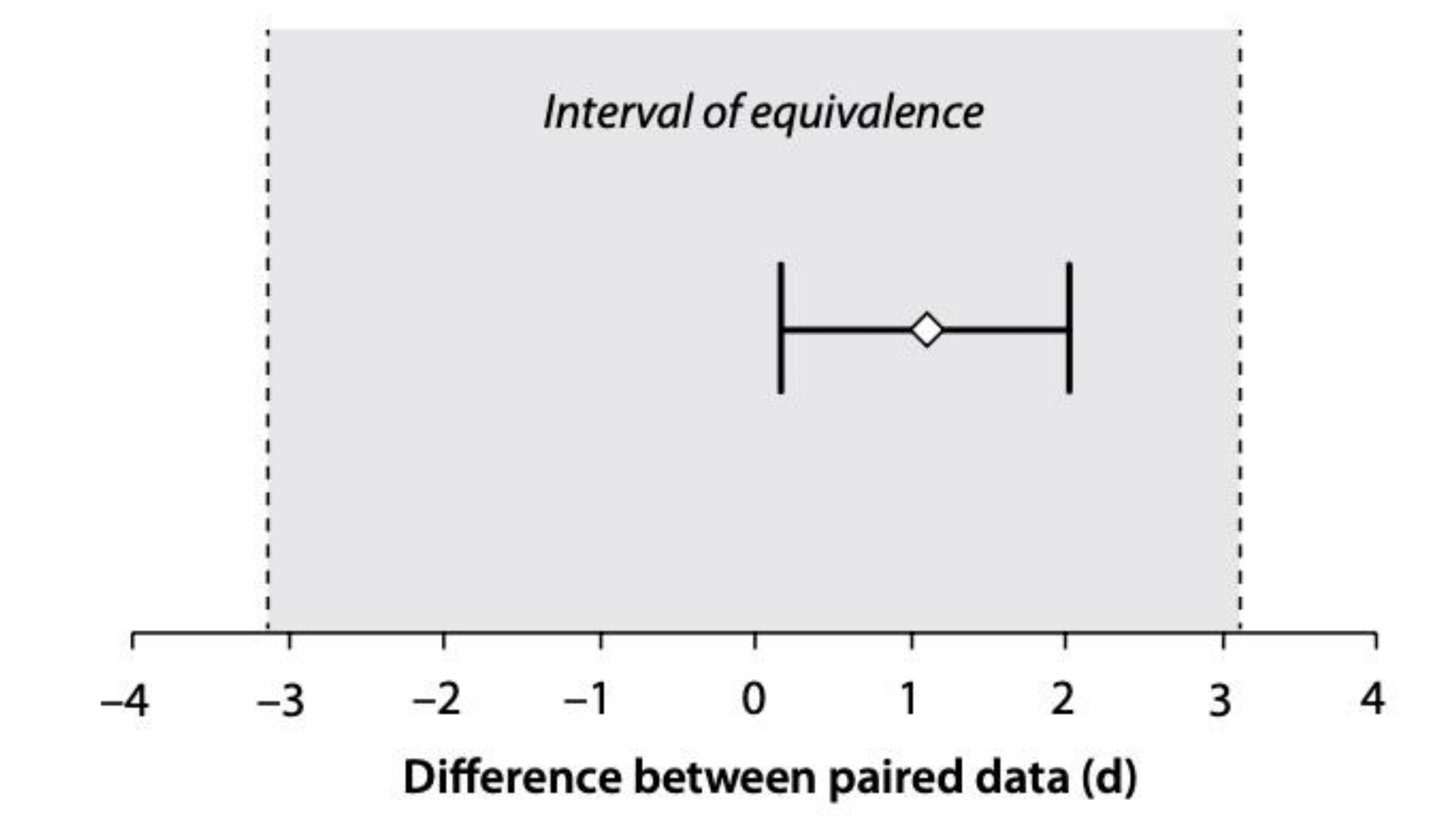
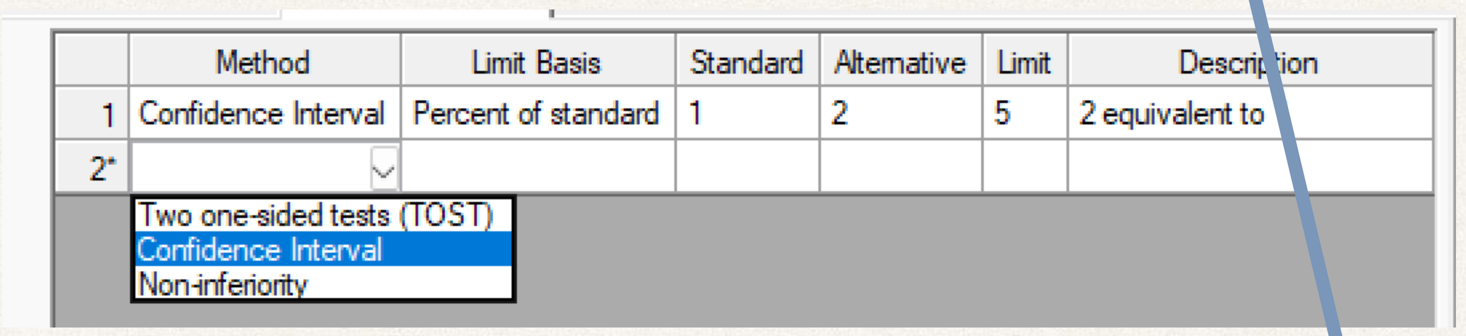
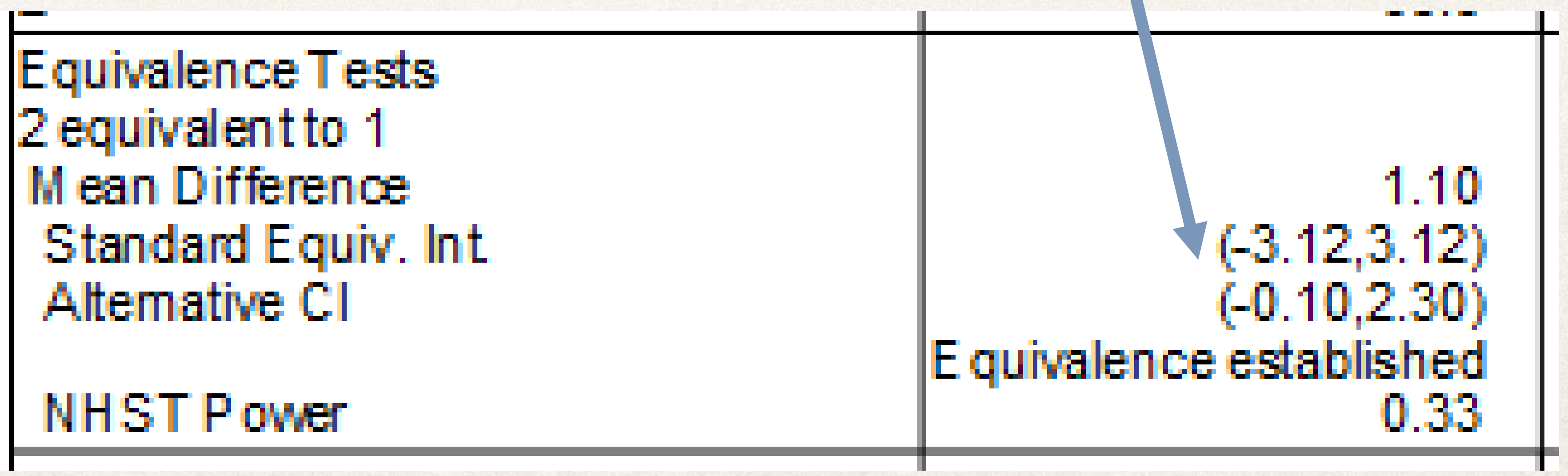


FIGURE 2. The confidence interval approach (Westlake's method) for TOST-P. The diamond represents the average difference (d = 1.1), while the whiskers are the 90% CI (0.18; 2.20); the grey shaded area is the interval of equivalence with the dashed lines marking its boundaries (-3.12; 3.12).



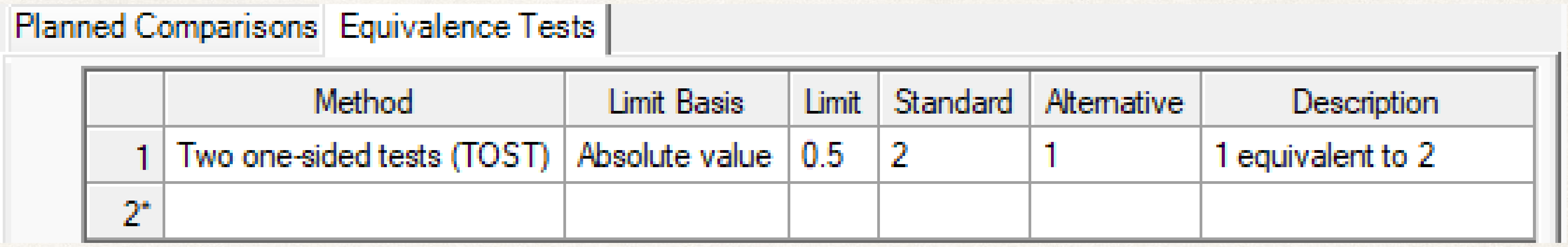
	Method	Limit Basis	Standard	Alternative	Limit	Description
1	Confidence Interval	Percent of standard	1	2	5	2 equivalent to
2*						



```
Equivalence Tests
2 equivalent to 1
 Mean Difference                         1.10
  Standard Equiv. Int            (-3.12,3.12)
  Alternative CI                 (-0.10,2.30)
                      Equivalence established
  NHST Power                             0.33
```

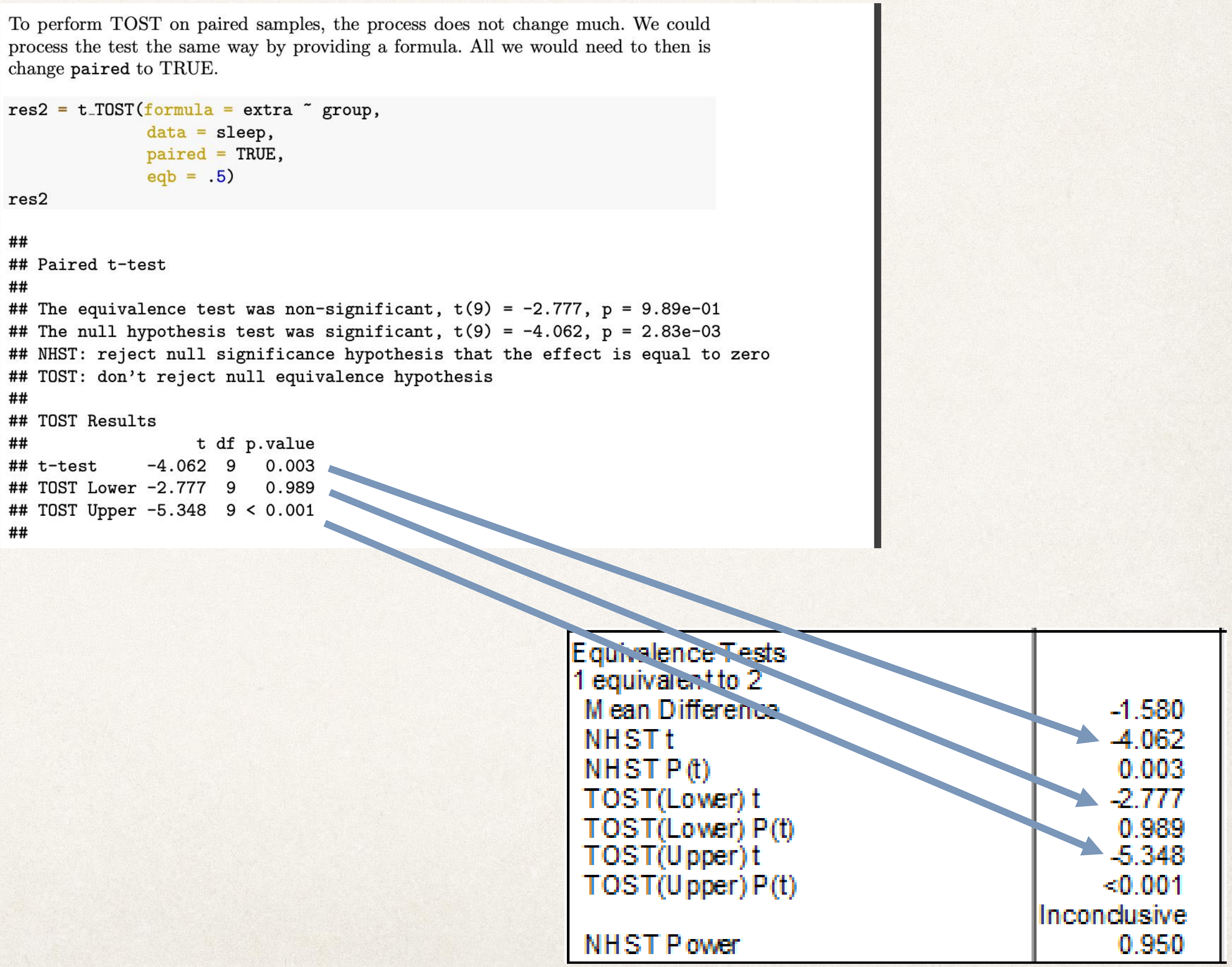
Caldwell 2022

- Caldwell contributed to the R TOSTER library, and in Caldwell 2022 provided examples of the use in this library.
- One data set was from a paired sample sleep study. To duplicate this in ARM, we use an RCB design, with pair as the blocking variable.
- Caldwell specifies an absolute value of 0.5 as the equivalence bound.
- To match TOST Lower and TOST Lower, we need to specify treatment 2 as the standard.
- A. R. Caldwell. Exploring equivalence testing with the updated TOSTER R package. PsyArXiv, November 2022.



Planned Comparisons | Equivalence Tests

	Method	Limit Basis	Limit	Standard	Alternative	Description
1	Two one-sided tests (TOST)	Absolute value	0.5	2	1	1 equivalent to 2
2*						



To perform TOST on paired samples, the process does not change much. We could process the test the same way by providing a formula. All we would need to then is change paired to TRUE.

```
res2 = t_TOST(formula = extra ~ group,
              data = sleep,
              paired = TRUE,
              eqb = .5)
res2
```

```
##
## Paired t-test
##
## The equivalence test was non-significant, t(9) = -2.777, p = 9.89e-01
## The null hypothesis test was significant, t(9) = -4.062, p = 2.83e-03
## NHST: reject null significance hypothesis that the effect is equal to zero
## TOST: don't reject null equivalence hypothesis
##
## TOST Results
##                  t df p.value
## t-test      -4.062  9   0.003
## TOST Lower  -2.777  9   0.989
## TOST Upper  -5.348  9 < 0.001
##
```

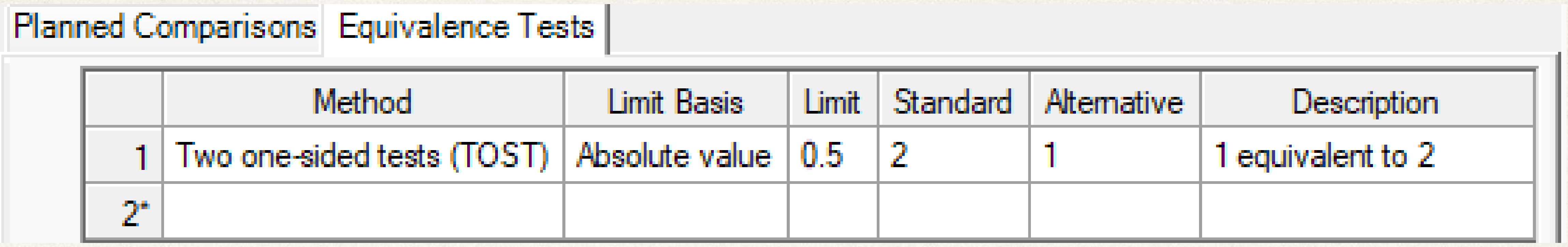
Equivalence Tests	
1 equivalent to 2	
Mean Difference	-1.580
NHST t	-4.062
NHST P(t)	0.003
TOST(Lower) t	-2.777
TOST(Lower) P(t)	0.989
TOST(Upper) t	-5.348
TOST(Upper) P(t)	<0.001
	Inconclusive
NHST Power	0.950

Caldwell 2022

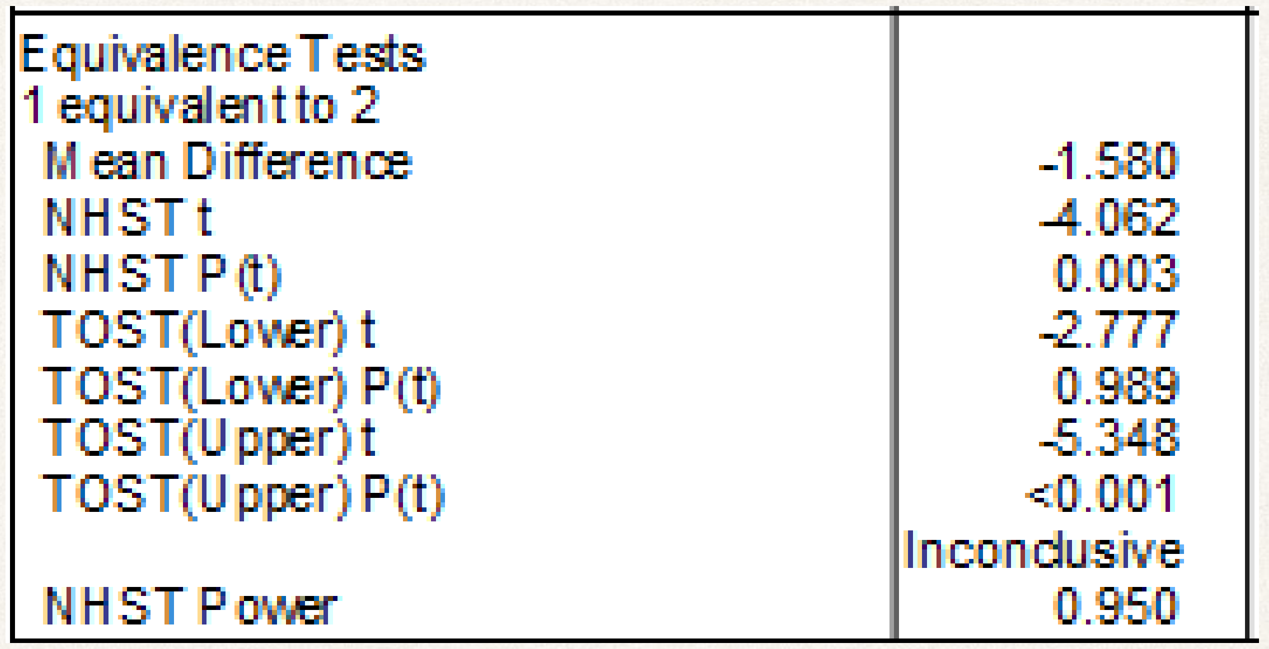
- We should note that the equivalence test is inconclusive (TOST Lower is not significant), and that the NHST Power is 0.950

- As we might expect from the NHST Power, we do indeed find a statistically significant difference between the two means.

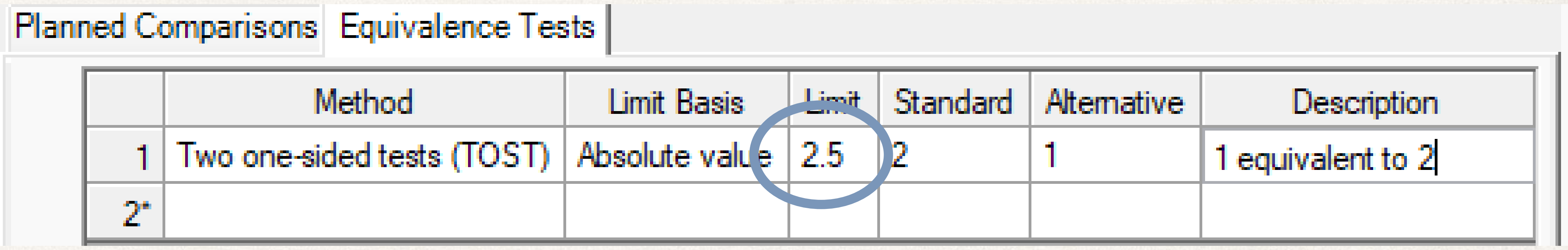
- We can produce the scenario where two treatments are both statistically significant and statistically equivalent if we increase the equivalence limit from 0.5 to 2.5. This illustrates the importance of a careful choice of equivalence bounds.



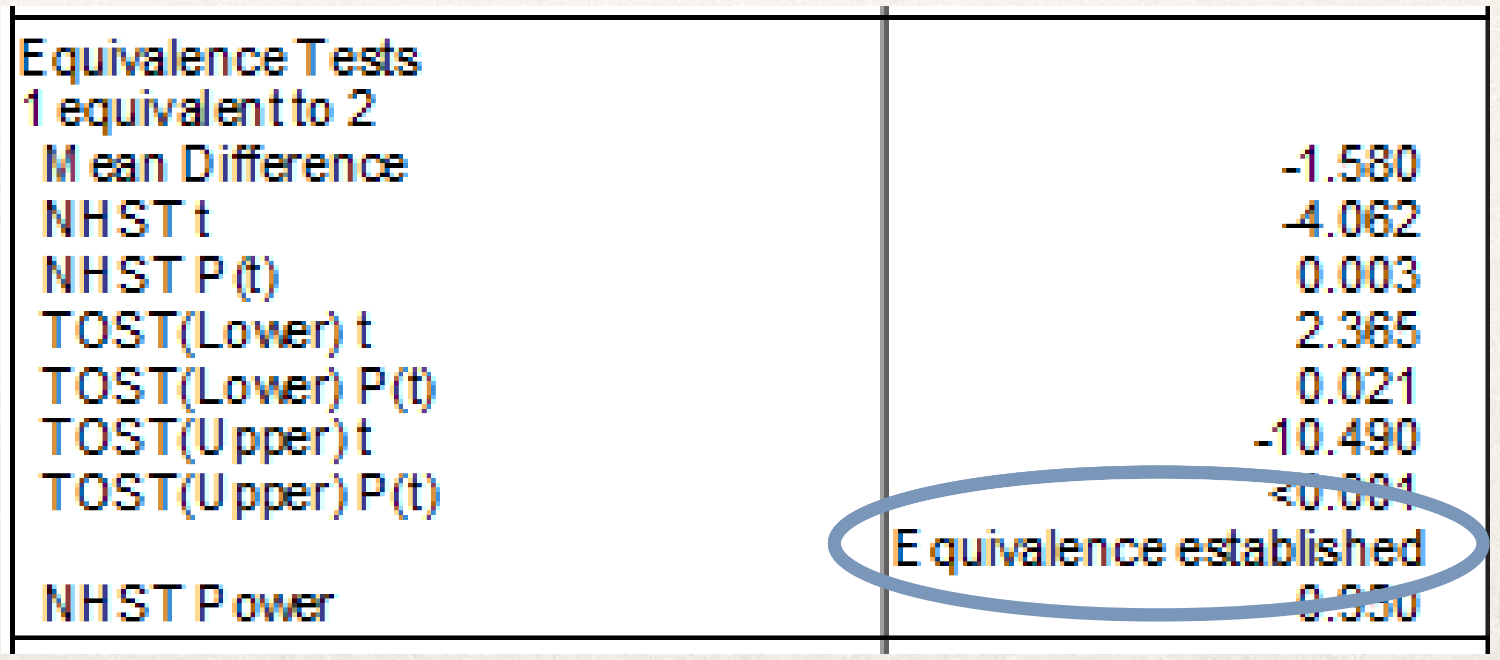
	Method	Limit Basis	Limit	Standard	Alternative	Description
1	Two one-sided tests (TOST)	Absolute value	0.5	2	1	1 equivalent to 2
2*						



```
Equivalence Tests
1 equivalent to 2
  Mean Difference                   -1.580
  NHST t                            -4.062
  NHST P(t)                          0.003
  TOST(Lower) t                     -2.777
  TOST(Lower) P(t)                   0.989
  TOST(Upper) t                     -5.348
  TOST(Upper) P(t)                  <0.001
                              Inconclusive
  NHST Power                         0.950
```

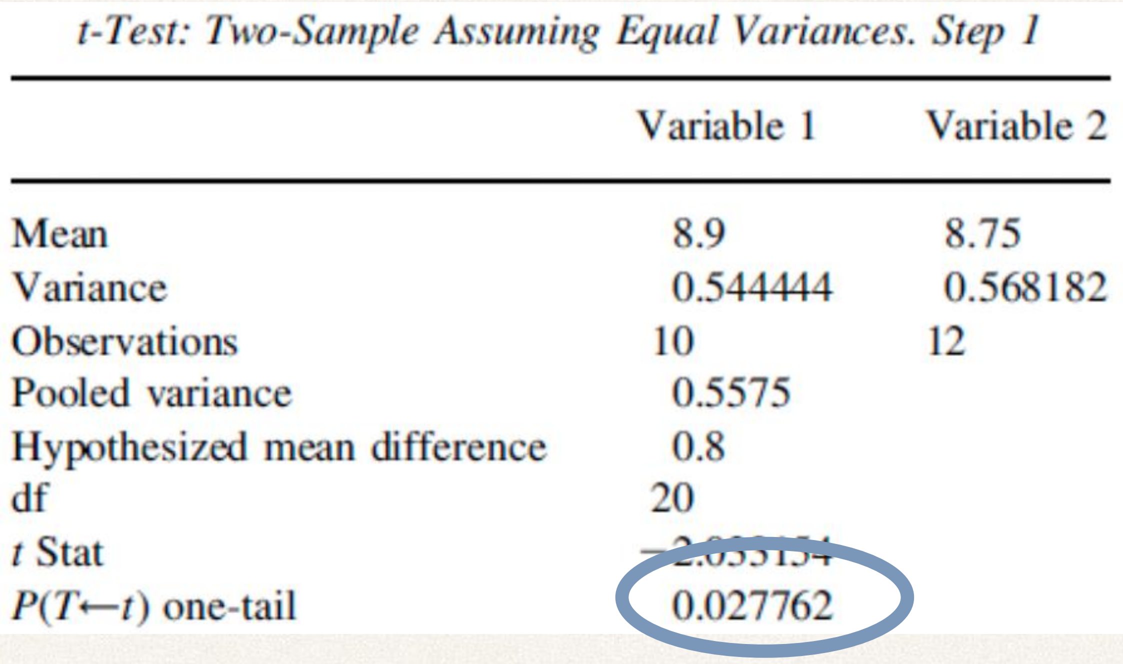


	Method	Limit Basis	Limit	Standard	Alternative	Description
1	Two one-sided tests (TOST)	Absolute value	2.5	2	1	1 equivalent to 2
2*						



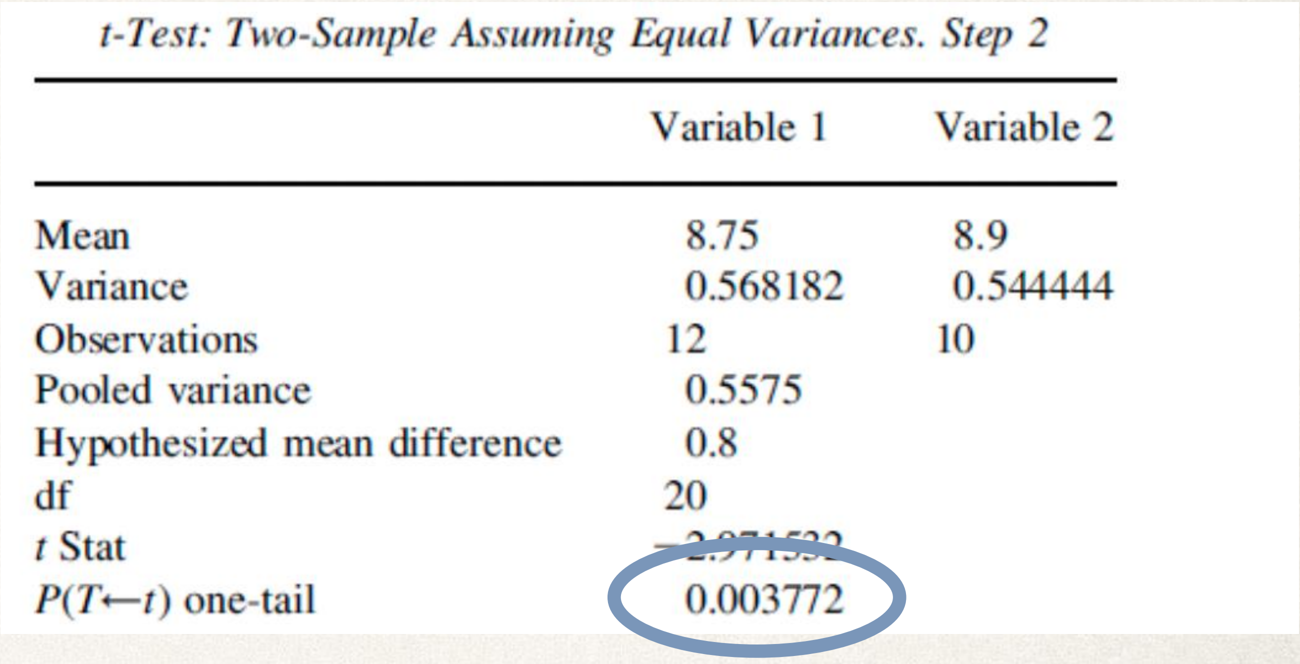
```
Equivalence Tests
1 equivalent to 2
  Mean Difference                   -1.580
  NHST t                            -4.062
  NHST P(t)                          0.003
  TOST(Lower) t                      2.365
  TOST(Lower) P(t)                   0.021
  TOST(Upper) t                    -10.490
  TOST(Upper) P(t)                  <0.001
                     Equivalence established
  NHST Power                         0.950
```

Richter 2002



t-Test: Two-Sample Assuming Equal Variances. Step 1

	Variable 1	Variable 2
Mean	8.9	8.75
Variance	0.544444	0.568182
Observations	10	12
Pooled variance	0.5575	
Hypothesized mean difference	0.8	
df	20	
t Stat	−2.033154	
P(*T*←*t*) one-tail	0.027762	



t-Test: Two-Sample Assuming Equal Variances. Step 2

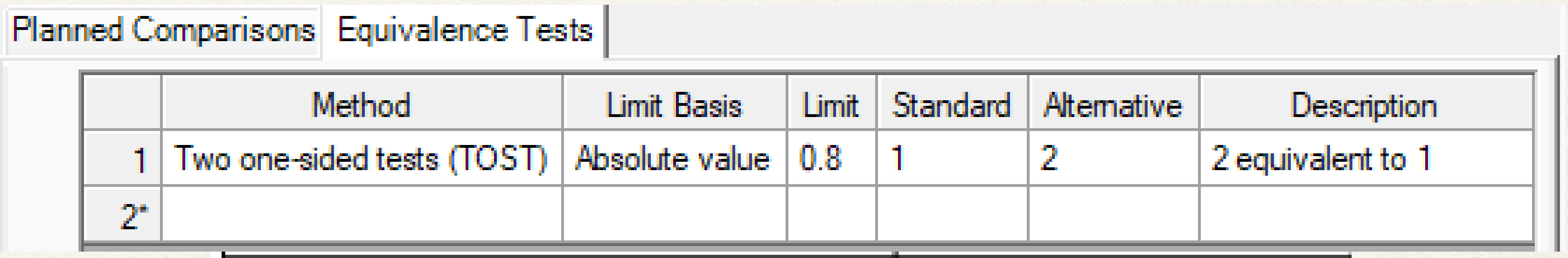
	Variable 1	Variable 2
Mean	8.75	8.9
Variance	0.568182	0.544444
Observations	12	10
Pooled variance	0.5575	
Hypothesized mean difference	0.8	
df	20	
t Stat	−2.971532	
P(*T*←*t*) one-tail	0.003772	

- Richter provides artificial data for a two sample independent means t-test. In ARM, we enter this as a CRD experiment.

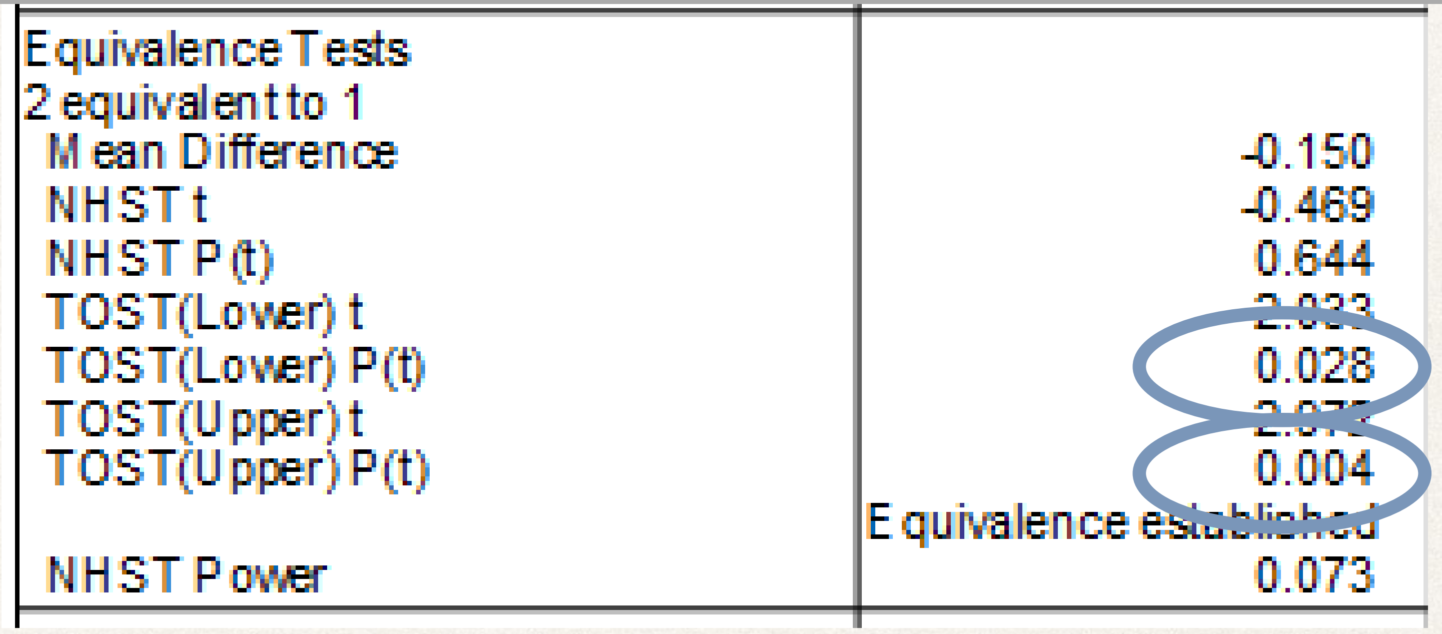
- Richter also provides instructions for entering the test data in an Excel spreadsheet. Importantly, the variables are reversed for step 2. This leads to a different sign for the t-test in step 2.

- In ARM, we don't reverse variables, so the signs of the two t-tests may be different. In either case, we compute comparable p-values for the t statistics.

- Since both TOST p-values are significant, we declare the two treatment groups equivalent within the 0.8 absolute value limit.



	Method	Limit Basis	Limit	Standard	Alternative	Description
1	Two one-sided tests (TOST)	Absolute value	0.8	1	2	2 equivalent to 1
2*						



```
Equivalence Tests
2 equivalent to 1
  Mean Difference                    -0.150
  NHST t                             -0.469
  NHST P(t)                           0.644
  TOST(Lower) t                       2.033
  TOST(Lower) P(t)                    0.028
  TOST(Upper) t                       2.972
  TOST(Upper) P(t)                    0.004
                     Equivalence established
  NHST Power                          0.073
```

- S. J. Richter and C. Richter. A method for determining equivalence in industrial applications. Quality Engineering, 14(3):375–380, 2002.

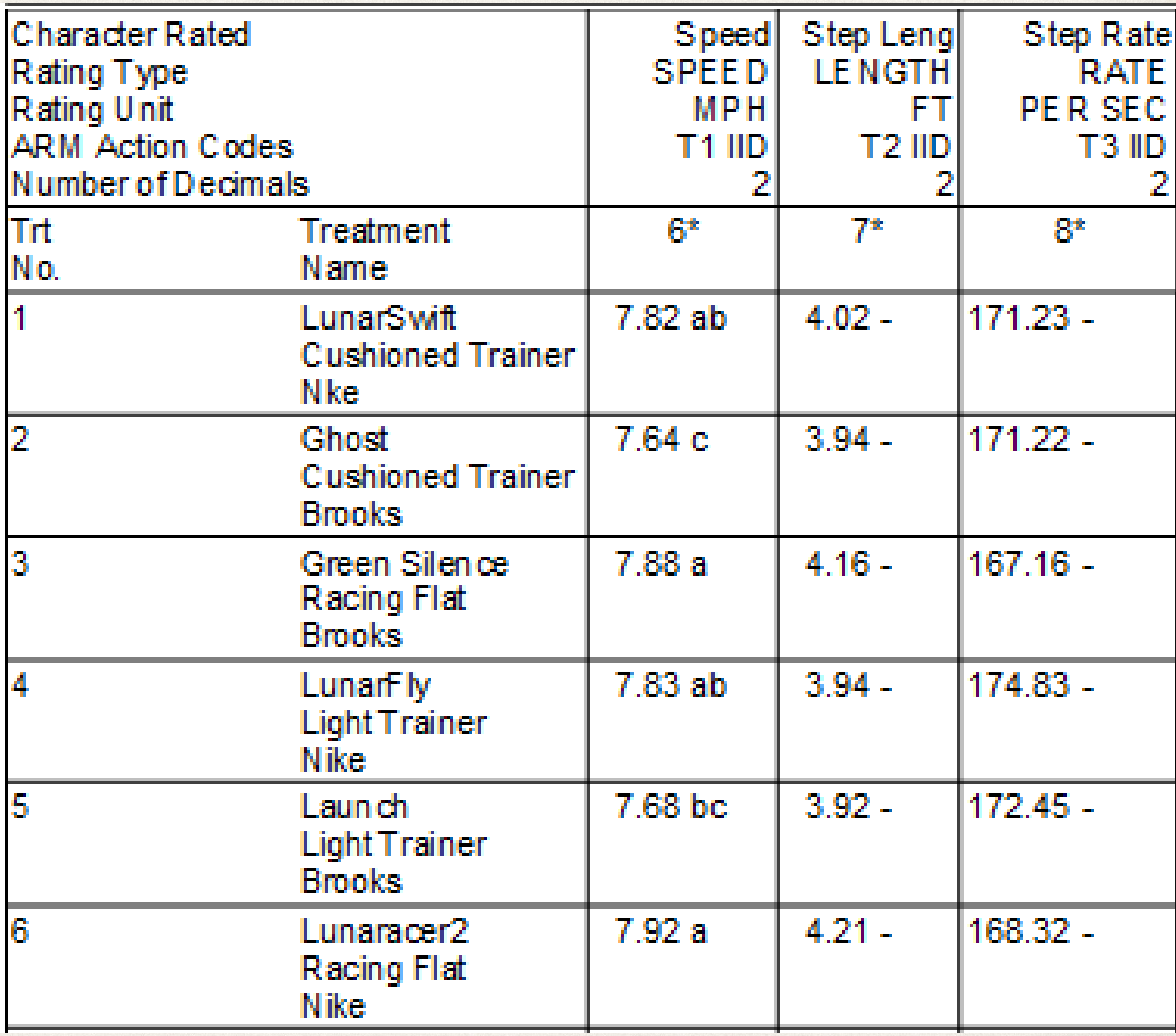
A more detailed example

- The previous examples from the literature focus on randomized control trials, which frequently have only two groups - treated and untreated, or treated with a standard treatment and treated with a new or alternative treatment.
- These examples also only focused on a single outcome.
- ARM trials are more commonly executed as blocked designs with multiple treatments, and usually involve multiple outcomes or assessments. We now consider

Shoe Trial

- I conducted a shoe trial where I ran in different models of running shoes, and I recorded as the assessments of interest running speed, step length and stride rate.

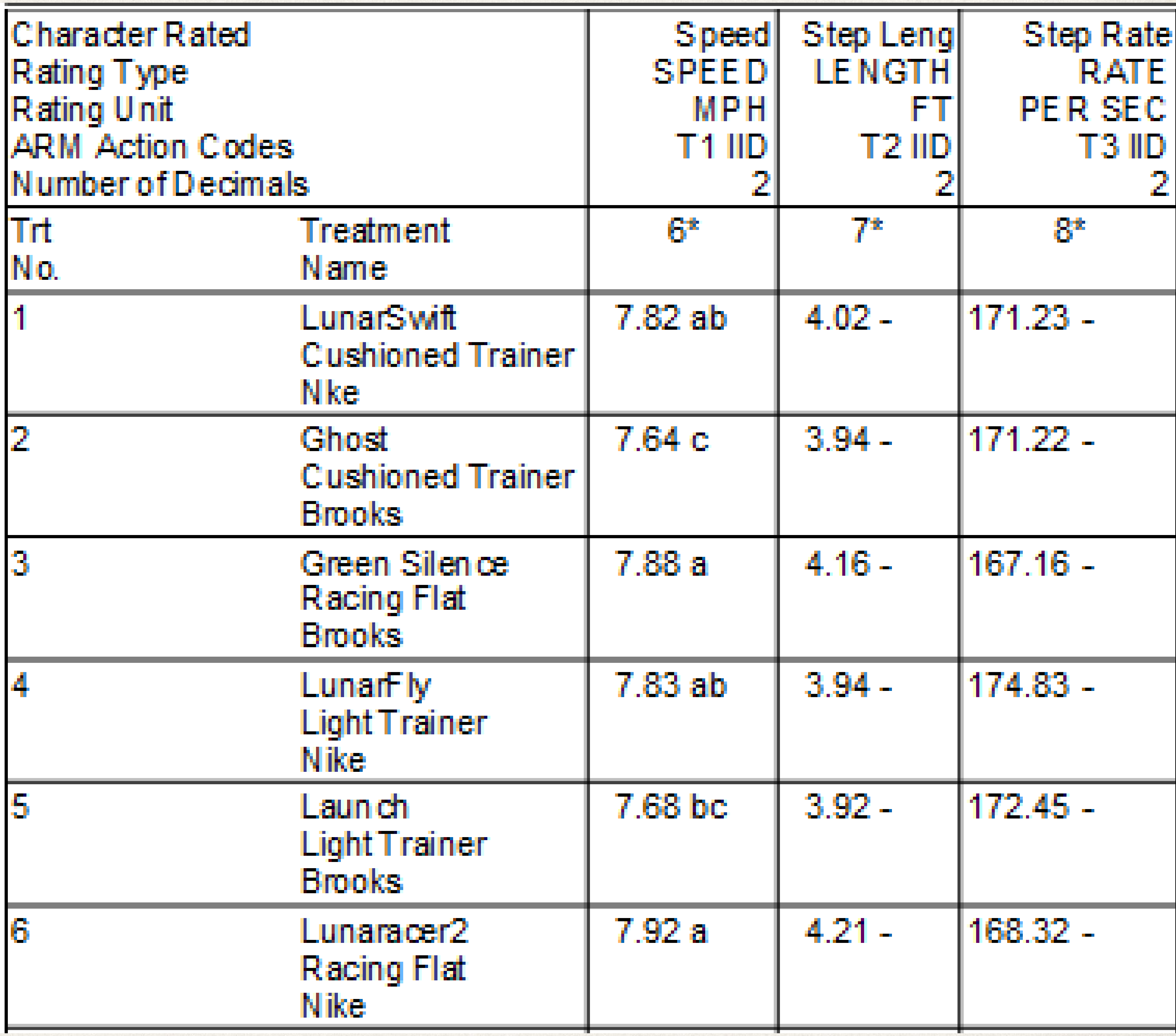
- Use the Fisher LSD test, I find that there are significant differences among running shoes - most notably, I ran faster (7.92 mph) in the racing flat Lunaracer2 than in the heavier training shoe Ghost (7.64 mph).



Character Rated		Speed	Step Leng	Step Rate
Rating Type		SPEED	LENGTH	RATE
Rating Unit		MPH	FT	PER SEC
ARM Action Codes		T1 IID	T2 IID	T3 IID
Number of Decimals		2	2	2
Trt No.	Treatment Name	6*	7*	8*
1	LunarSwift Cushioned Trainer Nke	7.82 ab	4.02 -	171.23 -
2	Ghost Cushioned Trainer Brooks	7.64 c	3.94 -	171.22 -
3	Green Silence Racing Flat Brooks	7.88 a	4.16 -	167.16 -
4	LunarFly Light Trainer Nike	7.83 ab	3.94 -	174.83 -
5	Launch Light Trainer Brooks	7.68 bc	3.92 -	172.45 -
6	Lunaracer2 Racing Flat Nike	7.92 a	4.21 -	168.32 -

Shoe Trial

- However, another model of shoe, Green Silence, is also a racing flat. Green Silence has the same mean separation letter (“a”) as Lunarracer2. That tells me the shoes are not statistically different.
- But are they equal - that is, does it matter if I want to run a fast race which pair of shoes I wear?
- I can’t statistically test that the shoes are equal, as we’ve discussed. But I can test if the shoes are equivalent.



Character Rated		Speed	Step Leng	Step Rate
Rating Type		SPEED	LENGTH	RATE
Rating Unit		MPH	FT	PER SEC
ARM Action Codes		T1 IID	T2 IID	T3 IID
Number of Decimals		2	2	2
Trt No.	Treatment Name	6*	7*	8*
1	LunarSwift Cushioned Trainer Nke	7.82 ab	4.02 -	171.23 -
2	Ghost Cushioned Trainer Brooks	7.64 c	3.94 -	171.22 -
3	Green Silence Racing Flat Brooks	7.88 a	4.16 -	167.16 -
4	LunarFly Light Trainer Nike	7.83 ab	3.94 -	174.83 -
5	Launch Light Trainer Brooks	7.68 bc	3.92 -	172.45 -
6	Lunaracer2 Racing Flat Nike	7.92 a	4.21 -	168.32 -

Shoe Trial

- Suppose I run in the Lunarracer a 5K in 20 minutes. A 1% difference in running speed is 12 seconds. A 5% difference in performance is 60 seconds - a full minute slower. I'm willing to accept two models of running shoe as equivalent if the difference between the two models is 1%; 12 seconds is acceptable. A 5% difference is too great to consider two models equivalent for racing purposes.

- So, I've determined that the smallest effect size of interest (SESOI), for the purpose of running speeds, to be 1% of the standard.

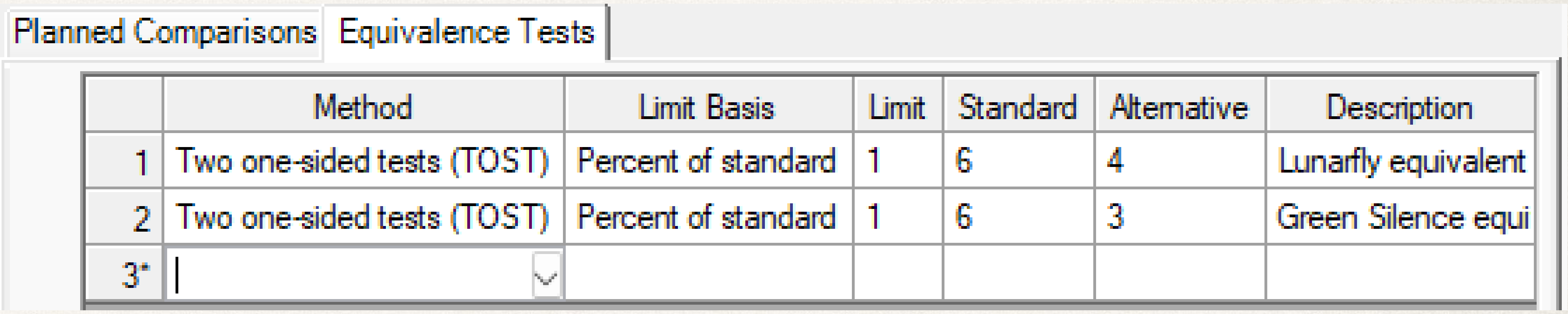
- Now I can ask the question -Are the two lightweight models of shoe, Lunarracer2 and Green Silence equivalent for my purposes?

- From the means table, I also note that a lightweight trainer, Lunarfly, has the same mean separation letter ("a") as Lunaracer2. The Lunarfly might be generally more comfortable and supportive, so I interested in determining if Lunaracer2 and Lunarfly are equivalent.

Shoe Trial

- Since Lunaracer2 (treatment 6) is common to both equivalence tests, we'll specify that model as the standard, and Lunarfly (treatment 4) and Green Silence (treatment 3) as alternatives.

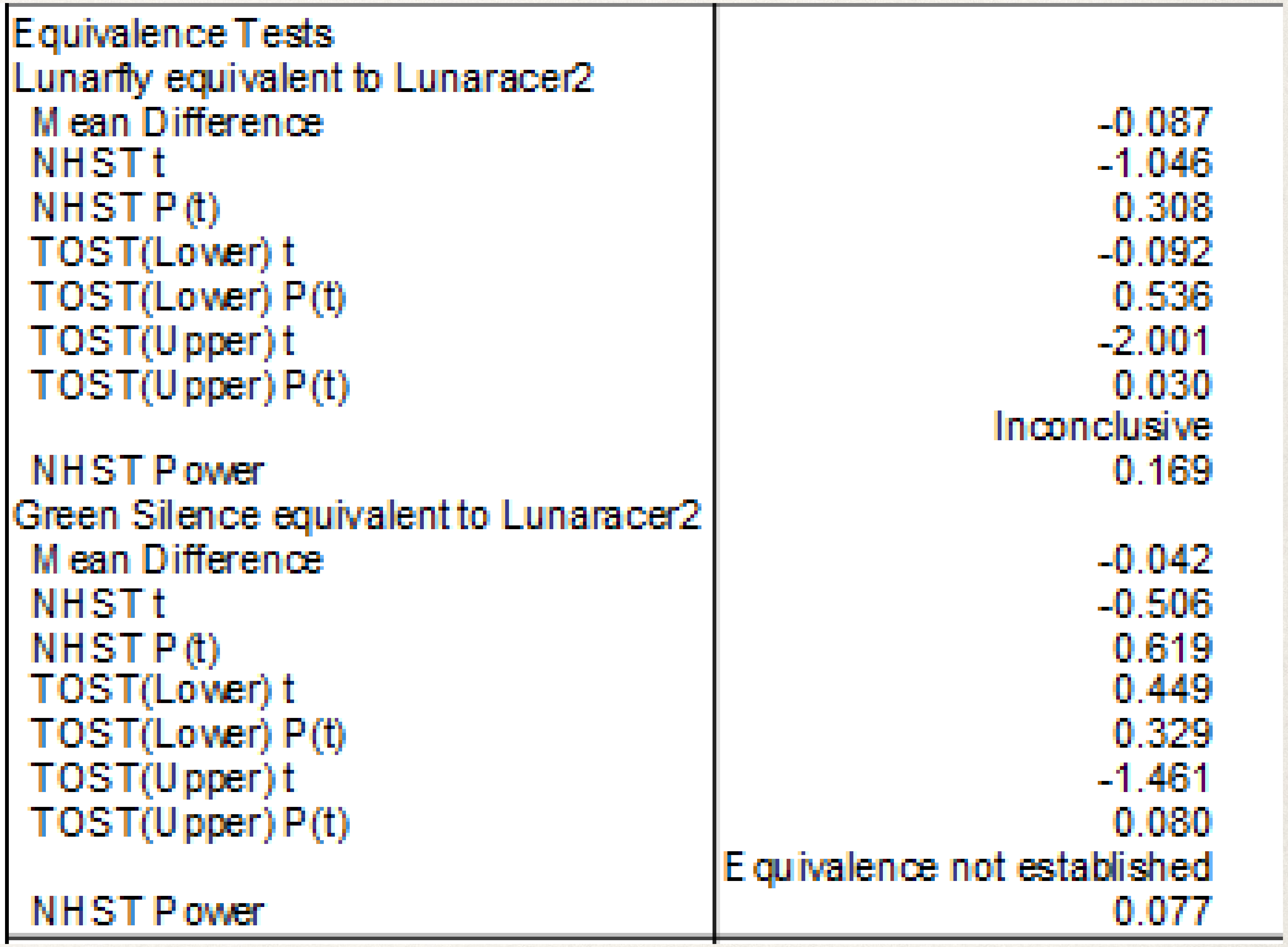
- The equivalence method is two one-sided tests, while the equivalence basis is 1% of standard (Lunaracer2 mean)



	Method	Limit Basis	Limit	Standard	Alternative	Description
1	Two one-sided tests (TOST)	Percent of standard	1	6	4	Lunarfly equivalent
2	Two one-sided tests (TOST)	Percent of standard	1	6	3	Green Silence equi
3*						

Shoe Trial

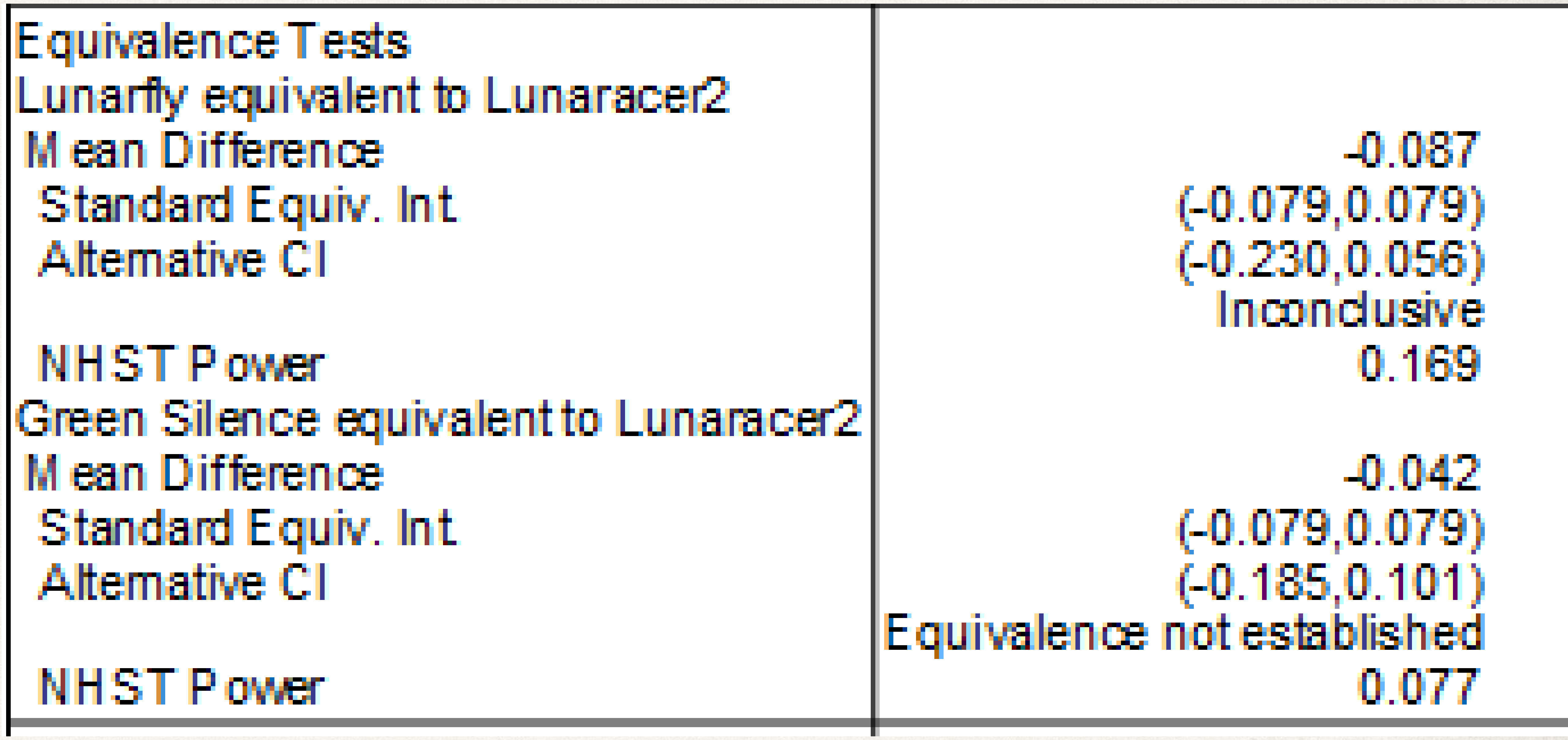
- We’ll focus on the Equivalence Tests for running speed only.
- We see that we cannot declare that Green Silence is equivalent to Lunaracer2, but the results for Lunarfly is inconclusive.
- This seems counter intuitive. The absolute value of mean difference for Lunarfly is larger than the difference for Green Silence.



Equivalence Tests	
Lunarfly equivalent to Lunaracer2	
Mean Difference	-0.087
NHST t	-1.046
NHST P(t)	0.308
TOST(Lower) t	-0.092
TOST(Lower) P(t)	0.536
TOST(Upper) t	-2.001
TOST(Upper) P(t)	0.030
	Inconclusive
NHST Power	0.169
Green Silence equivalent to Lunaracer2	
Mean Difference	-0.042
NHST t	-0.506
NHST P(t)	0.619
TOST(Lower) t	0.449
TOST(Lower) P(t)	0.329
TOST(Upper) t	-1.461
TOST(Upper) P(t)	0.080
	Equivalence not established
NHST Power	0.077

Shoe Trial

- Perhaps the CI method is more revealing.
- We see that for Lunarfly, the upper bound of the mean difference CI is contained within the equivalence interval. It is not likely that Lunarfly is superior to Lunaracer2, so is equivalent with respect to the upper bound. But it may also be a much slower shoe, so equivalence is inconclusive.
- In contrast, given the limits of equivalence, Green Silence may be a faster shoe, but may also be a slower shoe - the equivalence interval is completely contained in the mean different CI, so we fail to reject both null hypothesis about equivalence.

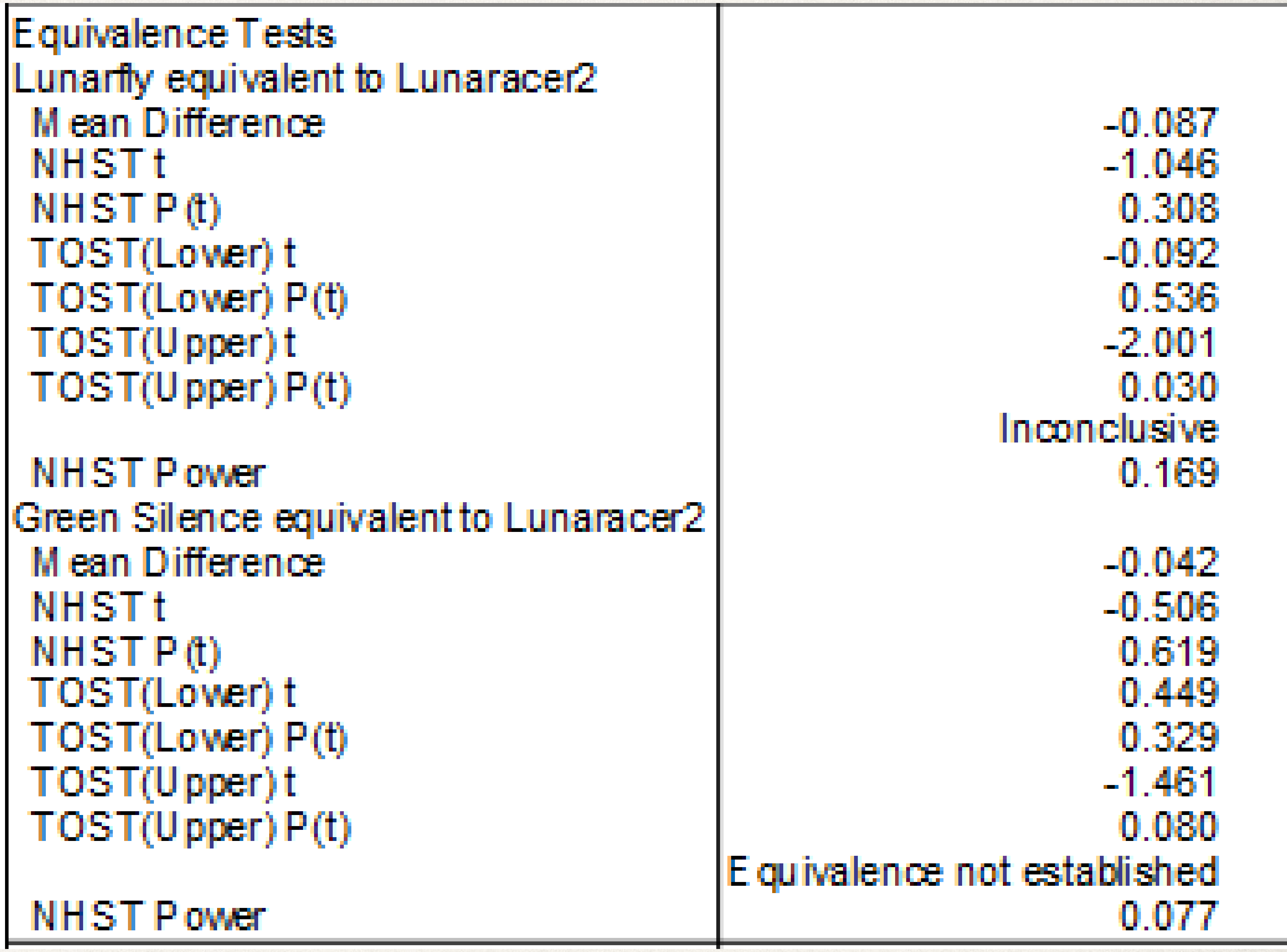


Equivalence Tests	
Lunarfly equivalent to Lunaracer2	
Mean Difference	-0.087
Standard Equiv. Int	(-0.079,0.079)
Alternative CI	(-0.230,0.056)
	Inconclusive
NHST Power	0.169
Green Silence equivalent to Lunaracer2	
Mean Difference	-0.042
Standard Equiv. Int	(-0.079,0.079)
Alternative CI	(-0.185,0.101)
	Equivalence not established
NHST Power	0.077

Shoe Trial

- If we refer back to Iolango, we may use the language that Lunarfly fails the null (one-sided) hypothesis of superiority, so we reject the null hypothesis of superiority, so we might conclude that Lunarfly is not superior to Luarracer2.

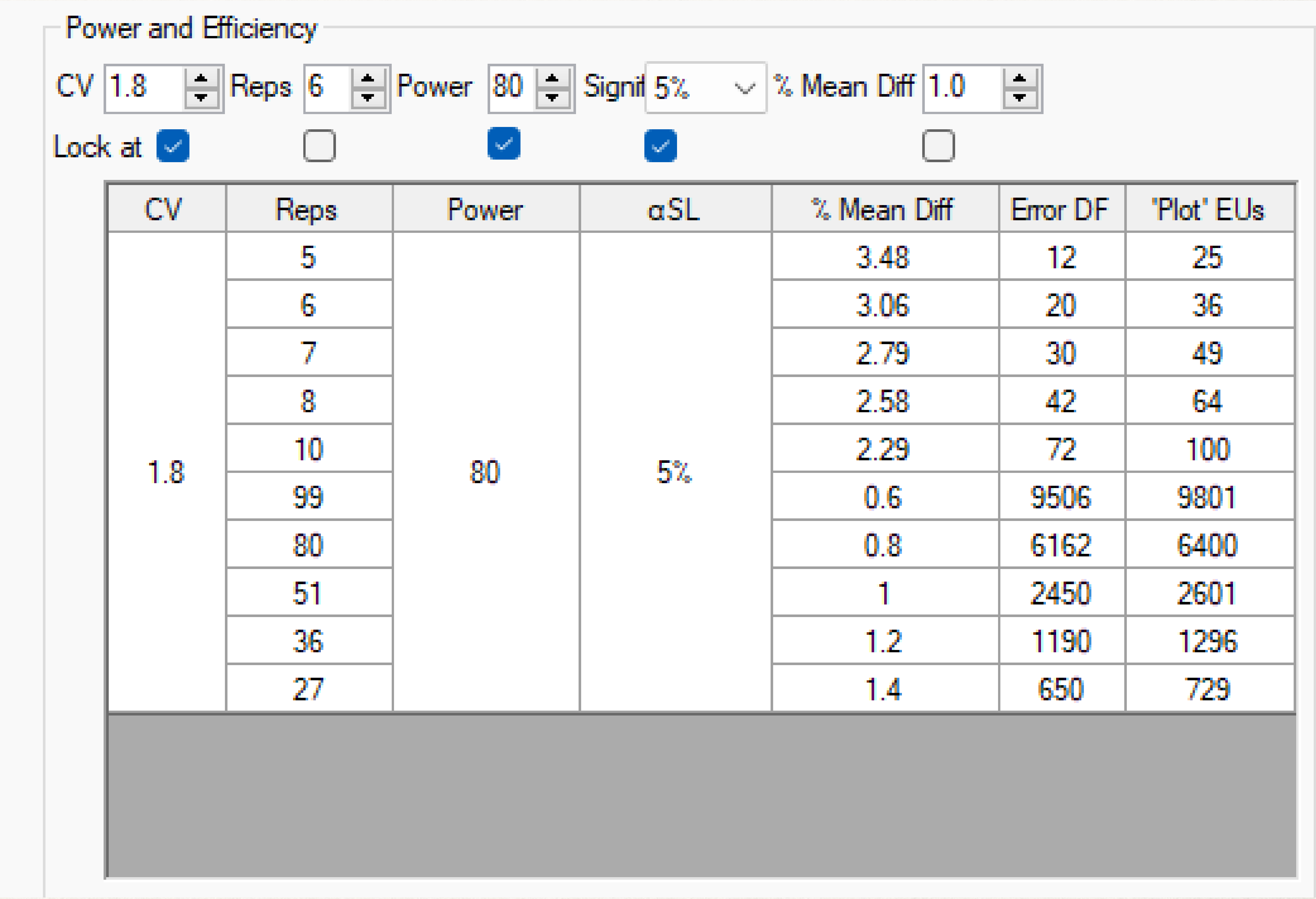
- Because we reject one and only one of the TOST, ARM reports this as “Inconclusive”. In contrast, we do not reject either null TOST hypothesis for Green Silence, we report the result that "Equivalence is not established”.



Equivalence Tests	
Lunarfly equivalent to Lunaracer2	
Mean Difference	-0.087
NHST t	-1.046
NHST P(t)	0.308
TOST(Lower) t	-0.092
TOST(Lower) P(t)	0.536
TOST(Upper) t	-2.001
TOST(Upper) P(t)	0.030
	Inconclusive
NHST Power	0.169
Green Silence equivalent to Lunaracer2	
Mean Difference	-0.042
NHST t	-0.506
NHST P(t)	0.619
TOST(Lower) t	0.449
TOST(Lower) P(t)	0.329
TOST(Upper) t	-1.461
TOST(Upper) P(t)	0.080
	Equivalence not established
NHST Power	0.077

Shoe Trial

- We should note that the computed statistical power of the NHST for these equivalence tests are low (0.169 and 0.077).
- If we entered the achieved CV in the Power and Efficiency table for this trial, we see that we would need 51 replicates to detect a mean difference of 1%, while with our given number of replicates (6) we could detect using the NHST a difference of 3%.



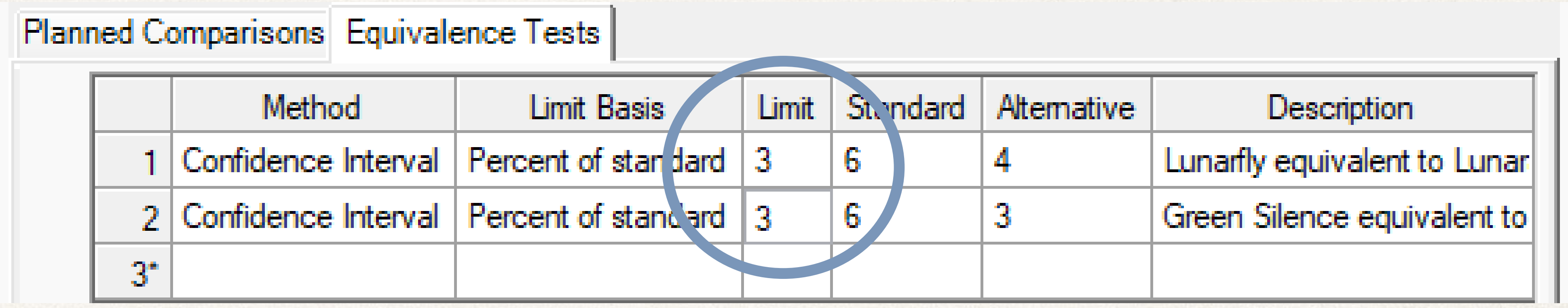
Power and Efficiency

CV 1.8 Reps 6 Power 80 Signi 5% % Mean Diff 1.0

Lock at ☑ □ ☑ ☑ □

CV	Reps	Power	αSL	% Mean Diff	Error DF	'Plot' EUs
1.8	5	80	5%	3.48	12	25
	6			3.06	20	36
	7			2.79	30	49
	8			2.58	42	64
	10			2.29	72	100
	99			0.6	9506	9801
	80			0.8	6162	6400
	51			1	2450	2601
	36			1.2	1190	1296
	27			1.4	650	729

Shoe Trial



Planned Comparisons | Equivalence Tests

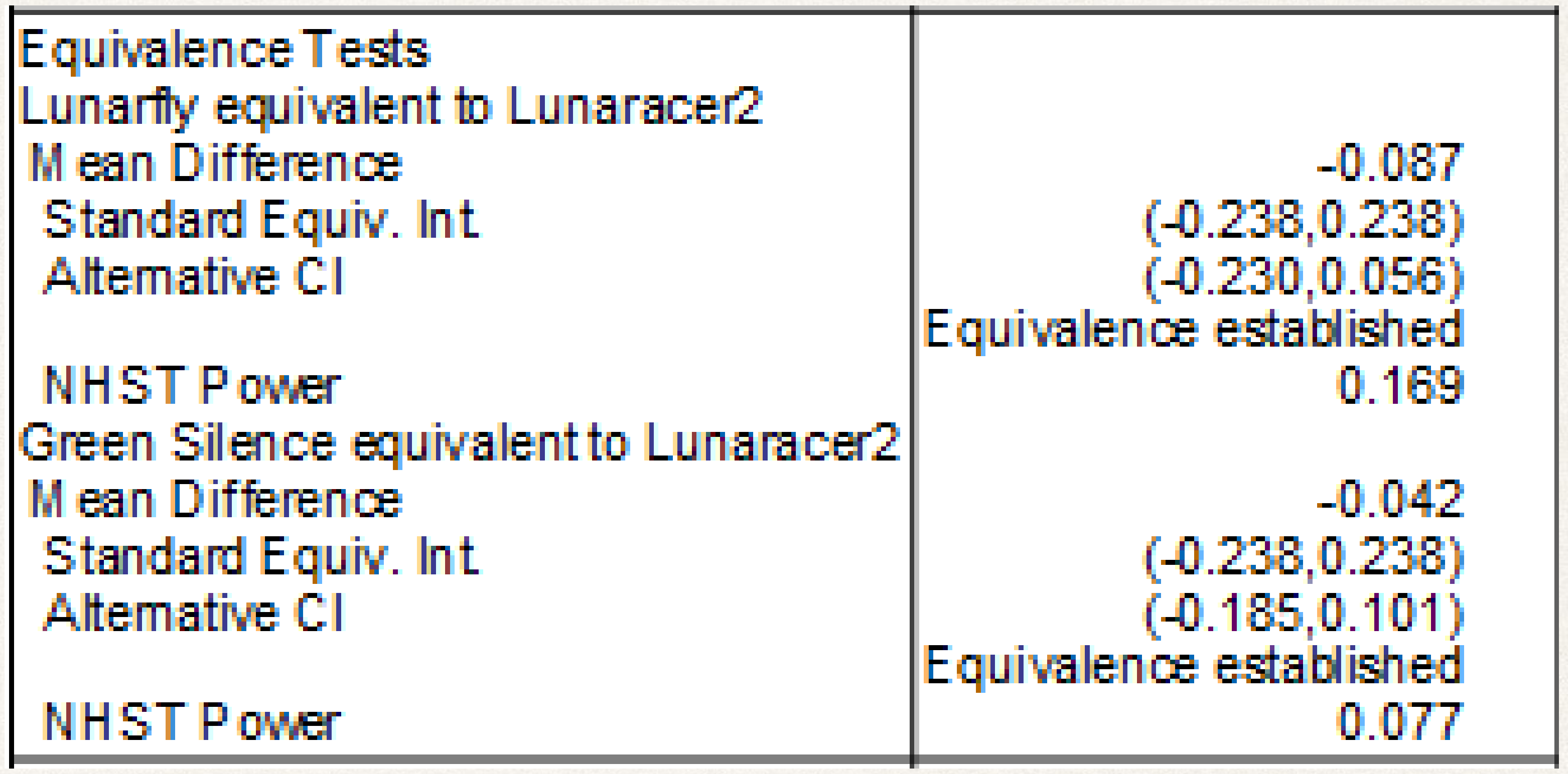
	Method	Limit Basis	Limit	Standard	Alternative	Description
1	Confidence Interval	Percent of standard	3	6	4	Lunarfly equivalent to Lunar
2	Confidence Interval	Percent of standard	3	6	3	Green Silence equivalent to
3*						

- Given that information, suppose we raise the equivalence limit from 1% to 3%.

- We then find equivalence between both pairs of running shoes.

- But is that good practice? Not really. We should be careful to specify a meaningful SESOI before we perform equivalence tests.

- Similarly, we should be careful in planning our experiments, so that we can detect differences close to our equivalence limits.



```
Equivalence Tests
Lunarfly equivalent to Lunaracer2
 Mean Difference                                   -0.087
  Standard Equiv. Int                      (-0.238,0.238)
  Alternative CI                           (-0.230,0.056)
                                   Equivalence established
  NHST Power                                        0.169
Green Silence equivalent to Lunaracer2
 Mean Difference                                   -0.042
  Standard Equiv. Int                      (-0.238,0.238)
  Alternative CI                           (-0.185,0.101)
                                   Equivalence established
  NHST Power                                        0.077
```

A bit more about the Equivalence limit

- Choosing the equivalent limit is not a statistical problem. Instead it requires, quoting Ialongo "demands the external support provided through the so-called equivalence interval to gain a meaning."

- As we've seen, it is possible to find a equivalent interval that supports equivalence for nearly any pair of means, if we first examine the data. For equivalence tests to be meaningful, they should be determined apriori.

A bit more about the Equivalence limit

- We also need to consider that equivalence limits may be determined by, again from Ialongo, different applications or different domains. For examples, the equivalence limit may be determine by assessment measurement; measurements taken with well-calibrate instruments may have narrower equivalence intervals than qualitative assessments.

- Due to the nature of ARM reports, our current design is limited to applying a single equivalence limit to the entire set of assessments included in reports. Thus, careful use of equivalence limits may require equivalence tests being reported only when a subset of assessments is chosen for a report.

A bit more about the Equivalence limit

- One method that may allow equivalence limits to be applied across multiple assessments in a single report is to use Cohen's d to define the equivalence limit.

- We not discussed the use of Cohen's d in this presentation since there are few examples in the literature that explicitly use d in defining equivalence limits. In this presentation, we've strived to show how ARM reproduces equivalence tests as reported in the literature.

- Cohen's d standardizes an effect to make the effect relative to variance of the data; d is usually given by $d = \left(\overline{X}_{new} - \overline{X}_{std}\right)/s$.

- However, thinking about mean and mean differences in terms of Cohen's d is not common in agronomic research and may require some adjustments to thinking about experimental results in agronomic sciences.

Conclusion

- Equivalence tests provide us a method of determining if treatments are effectively the same, when differences cannot be detected statistically using the traditional null hypothesis testing.
- However, we must take care that our experiments have sufficient power to detect important differences, and we must take care in determining SESOI.
- Determining SESOI is frequently not a statistical problem, but instead must take into consideration factors of treatment application and experimental outcome.