**Appendix 1, Supplementary online material.**

Conducting the Equivalence test in different statistical packages. In these examples, equivalence at 5 and 15µg are shown to demonstrate the contrast output when the methods are determined to be equivalent and when they are not equivalent. The dataset is also available on request (marijka@uow.edu.au) to replicate the analyses.

1. R

In R equivalence testing can be conducted easily using the package “equivalence” ([www.cran.**r**-project.org/web/**packages**/**equivalence**/**equivalence**.pdf](http://www.cran.r-project.org/web/packages/equivalence/equivalence.pdf))

The tost command (in bold) can be used for paired or independent data, by specifying a single variable the paired test is used. The test is conducted on the bias, the difference between the methods. In this analysis this variable is called “bias” and the dataset is called “iodine”. The bias is the difference between the x24hrR (average of 3 24 hour recalls) and the FFQ (food frequency questionnaire)

**tost(iodine$bias, y=NULL, alpha=0.05, epsilon=5)**

$mean.diff

[1] 2.284939

$se.diff

[1] 4.91133

$alpha

[1] 0.05

$ci.diff

[1] -5.889338 10.459215

attr(,"conf.level")

[1] 0.9

$df

df

79

$epsilon

[1] 5

$result

[1] "not rejected"

$p.value

[1] 0.2909752

$check.me

[1] -0.430123 5.000000

attr(,"conf.level")

[1] 0.4180496

*The P value is >0.05 (0.2909752) and indicates the methods are not equivalent.*

**tost(iodine$bias, y=NULL, alpha=0.05, epsilon=15)**

$mean.diff

[1] 2.284939

$se.diff

[1] 4.91133

$alpha

[1] 0.05

$ci.diff

[1] -5.889338 10.459215

attr(,"conf.level")

[1] 0.9

$df

df

79

$epsilon

[1] 15

$result

[1] "rejected"

$p.value

[1] 0.00572852

$check.me

[1] -10.43012 15.00000

attr(,"conf.level")

[1] 0.988543

*The P value is <0.05 (0.00572852) and indicates the methods are equivalent.*

1. In SAS V9.3 (SAS Inc, Cary NC), equivalence testing is available through the PROC TTEST procedure, in SAS it is necessary to specify that the test is paired and the values for the FR and FFQ are used. SAS also produces graphical output. As mentioned in the text, the SAS macro concord(1) produces equivalence tests and Philip Dixon(2) provides syntax on determining equivalence using the PROC MIXED procedure in an online archive EquivSlope.sas <http://www.esapubs.org/archive/ecol/E086/094/suppl-1.htm>

**Proc** **ttest** data=iodine tost(-**5**,**5**);

paired FR\*FFQ;

**run**;

|  |
| --- |
| The SAS System |

The TTEST Procedure

Difference: FR - FFQ

| **N** | **Mean** | **Std Dev** | **Std Err** | **Minimum** | **Maximum** |
| --- | --- | --- | --- | --- | --- |
| 80 | -2.2849 | 43.9283 | 4.9113 | -119.8 | 76.2301 |

| **Mean** | **95% CL Mean** | **Std Dev** | **95% CL Std Dev** |
| --- | --- | --- | --- |
| -2.2849 | -12.0607 | 7.4908 | 43.9283 | 38.0178 | 52.0318 |

TOST Level 0.05 Equivalence Analysis

| **Mean** | **Lower Bound** |  | **90% CL Mean** |  | **Upper Bound** | **Assessment** |
| --- | --- | --- | --- | --- | --- | --- |
| -2.2849 | -5 | > | -10.4592 | 5.8893 | > | 5 | Not equivalent |

| **Test** | **Null** | **DF** | **t Value** | **P-Value** |
| --- | --- | --- | --- | --- |
| **Upper** | -5 | 79 | 0.55 | 0.2910 |
| **Lower** | 5 | 79 | -1.48 | 0.0710 |
| **Overall** |   |   |   | 0.2910 |

*The P value is >0.05 (0.2910) and indicates the methods are not equivalent.*



**Proc** **ttest** data=iodine tost(-**15**,**15**);

paired FR\*FFQ;

**run**;

|  |
| --- |
| The SAS System |

The TTEST Procedure

Difference: FR - FFQ

| **N** | **Mean** | **Std Dev** | **Std Err** | **Minimum** | **Maximum** |
| --- | --- | --- | --- | --- | --- |
| 80 | -2.2849 | 43.9283 | 4.9113 | -119.8 | 76.2301 |

| **Mean** | **95% CL Mean** | **Std Dev** | **95% CL Std Dev** |
| --- | --- | --- | --- |
| -2.2849 | -12.0607 | 7.4908 | 43.9283 | 38.0178 | 52.0318 |

TOST Level 0.05 Equivalence Analysis

| **Mean** | **Lower Bound** |  | **90% CL Mean** |  | **Upper Bound** | **Assessment** |
| --- | --- | --- | --- | --- | --- | --- |
| -2.2849 | -15 | < | -10.4592 | 5.8893 | < | 15 | Equivalent |

| **Test** | **Null** | **DF** | **t Value** | **P-Value** |
| --- | --- | --- | --- | --- |
| **Upper** | -15 | 79 | 2.59 | 0.0057 |
| **Lower** | 15 | 79 | -3.52 | 0.0004 |
| **Overall** |   |   |   | 0.0057 |

*The overall P value is less than 0.05 and indicates the methods are equivalent*



1. In STATA V12 (StataCorp LP, College Station, TX) equivalence tests are available through a user written .ado file written by Alexis Dinno available from <http://doyenne.com/stata/tost.html>

**tostt fr==ffq, eqvt(delta) eqvl(5)**

Paired t test of mean equivalence

------------------------------------------------------------------------------

Variable | Obs Mean Std. Err. Std. Dev. [95% Conf. Interval]

---------+--------------------------------------------------------------------

 fr | 80 124.2285 5.43549 48.6165 113.4094 135.0476

 ffq | 80 126.5134 6.043544 54.0551 114.484 138.5428

---------+--------------------------------------------------------------------

 D-diff | 7.284939 4.91133 -2.490819 17.0607

 diff+D | 2.715061 4.91133 -7.060696 12.49082

------------------------------------------------------------------------------

mean(diff) = mean(fr - ffq)

 Delta (D) = 5.0000 Delta expressed in same units as fr

Impossible to reject any Ho if Delta <= t-crit\*s.e. ( 8.174 ). See help tostt.

 df = 79

Ho: |diff| >= Delta:

 t1 = 1.483 t2 = .5528

 Ho1: Delta-diff >= 0 Ho2: diff+Delta <= 0

 Ha1: Delta-diff < 0 Ha2: diff+Delta > 0

 **Pr(T > t1) = 0.0710 Pr(T > t2) = 0.2910**

*Both P values must be significant for the methods to be equivalent, therefore not equivalent*

**tostt fr==ffq, eqvt(delta) eqvl(15)**

Paired t test of mean equivalence

------------------------------------------------------------------------------

Variable | Obs Mean Std. Err. Std. Dev. [95% Conf. Interval]

---------+--------------------------------------------------------------------

 fr | 80 124.2285 5.43549 48.6165 113.4094 135.0476

 ffq | 80 126.5134 6.043544 54.0551 114.484 138.5428

---------+--------------------------------------------------------------------

 D-diff | 17.28494 4.91133 7.509181 27.0607

 diff+D | 12.71506 4.91133 2.939304 22.49082

------------------------------------------------------------------------------

mean(diff) = mean(fr - ffq)

 Delta (D) = 15.0000 Delta expressed in same units as fr

 df = 79

Ho: |diff| >= Delta:

 t1 = 3.519 t2 = 2.589

 Ho1: Delta-diff >= 0 Ho2: diff+Delta <= 0

 Ha1: Delta-diff < 0 Ha2: diff+Delta > 0

 Pr(T > t1) = 0.0004 Pr(T > t2) = 0.0057

*Both P values must be significant for the methods to be equivalent, therefore equivalent*

1. In SPSS V21 (IBM Corporation, Armonk NY) there is not an automated procedure to produce the two one sided tests. This can be done manually by conducting two one sample t tests using the upper and lower equivalence values and the bias as the test variable. This test only produces a two tailed output of significance which needs to be halved for the one tailed P value. If both of these are significant then the methods are equivalent. This can be demonstrated by comparing with the STATA output above. Note that halving the P values is approximate, exact one sided P values could be obtained from several free online calculators or by using R (for example 1-pt(3.519, df=80) returns P=0.003588676.

T-TEST

 /TESTVAL=-5

 /MISSING=ANALYSIS

 /VARIABLES=bias

 /CRITERIA=CI(.95).

T-TEST

 /TESTVAL=5

 /MISSING=ANALYSIS

 /VARIABLES=bias

 /CRITERIA=CI(.95).

|  |
| --- |
| **One-Sample Test** |
|  | Test Value = -5 |
| t | df | Sig. (2-tailed) | Mean Difference | 95% Confidence Interval of the Difference |
| Lower | Upper |
| Bias | 1.483 | 79 | .142 | 7.28494 | -2.4908 | 17.0607 |

*The P value of 0.142 must be halved to give 0.071 for the lower equivalence bound*

|  |
| --- |
| **One-Sample Test** |
|  | Test Value = 5 |
| t | df | Sig. (2-tailed) | Mean Difference | 95% Confidence Interval of the Difference |
| Lower | Upper |
| Bias | -.553 | 79 | .582 | -2.71506 | -12.4908 | 7.0607 |

*The P value of 0.582 must be halved to give 0.291 for the upper equivalence bound. As neither of these are significant at the 0.05 level, the methods are not equivalent.*

T-TEST

 /TESTVAL=-15

 /MISSING=ANALYSIS

 /VARIABLES=bias

 /CRITERIA=CI(.95).

T-TEST

 /TESTVAL=15

 /MISSING=ANALYSIS

 /VARIABLES=bias

 /CRITERIA=CI(.95).

|  |
| --- |
| **One-Sample Test** |
|  | Test Value = -15 |
| t | df | Sig. (2-tailed) | Mean Difference | 95% Confidence Interval of the Difference |
| Lower | Upper |
| Bias | 3.519 | 79 | .001 | 17.28494 | 7.5092 | 27.0607 |

*The P value of 0.001 must be halved to give P=0.0005 for the lower equivalence bound*

|  |
| --- |
| **One-Sample Test** |
|  | Test Value = 15 |
| t | df | Sig. (2-tailed) | Mean Difference | 95% Confidence Interval of the Difference |
| Lower | Upper |
| Bias | -2.589 | 79 | .011 | -12.71506 | -22.4908 | -2.9393 |

*The P value of 0.011 must be halved to give P=0.006 for the upper equivalence bound. As BOTH of these tests are significant at the 0.05 level, the methods are equivalent.*

The SPSS custom dialog box SPSS custom dialog developed by Weber & Popova(3) available from <http://www.medianeuroscience.org/equivalence_testing> uses effect sizes based on Cohen’s d. In order to replicate the examples in this paper the equivalence bounds were converted to approximated effect sizes(4) for the upper and lower bound using the pooled standard deviation of the difference between the methods and correlation from the paired t test and then averaged to create an overall effect size for the 5 and 15 equivalent ranges. The default values for Cohen’s small, medium and large effect sizes are also presented as an alternative approach.

*For the equivalence bounds of (-5,5) the approximated effect size is 0.1328*

\*\*\* Weber & Popova Dependent/Paired-Samples Equivalence Procedure \*\*\*

 Based on the custom-entered delta

 p based on p based on

 actual value of delta half variance explained

Custom delta t df (two-tailed) (two-tailed)

\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

 .133 -.47 79 .188 .382

*For the equivalence bounds of (-15,15) the approximated effect size is 0.3734*

\*\*\* Weber & Popova Dependent/Paired-Samples Equivalence Procedure \*\*\*

 Based on the custom-entered delta

 p based on p based on

 actual value of delta half variance explained

Custom delta t df (two-tailed) (two-tailed)

\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

 .373 -.47 79 .000 .038

Using the default Cohen’s effect sizes.

\*\*\* Weber & Popova Dependent/Paired-Samples Equivalence Procedure \*\*\*

 Based on the Cohen's classification of effect sizes

 t df Delta p, two-tailed

\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_

 -.47 79 .10 .291

 -.47 79 .30 .004

 -.47 79 .50 .000

1. Groeneveld J (2011) Embedding equivalence t-test results in Bland Altman Plots visualising rater reliability. In *Pharmaceutical Users Software Exchange*, pp. SP06. Brighton, UK: PhUSE.

2. Dixon PM, Pechmann JHK (2005) A statistical test to show negligble trend. *Ecology* **86**, 1751-1756.

3. Weber R, Popova L (2012) Testing equivalence in communication research: Theory and application. *Communication methods and measures* **6**, 190-213.

4. Dunlap WP, Cortina JM, Vaslow JB *et al.* (1996) Meta-analysis of experiments with matched groups of repeated measures designs. *Psychological Methods* **1**, 170-177.