Dependent t-test

Daniel Lakens, D.Lakens@tue.nl

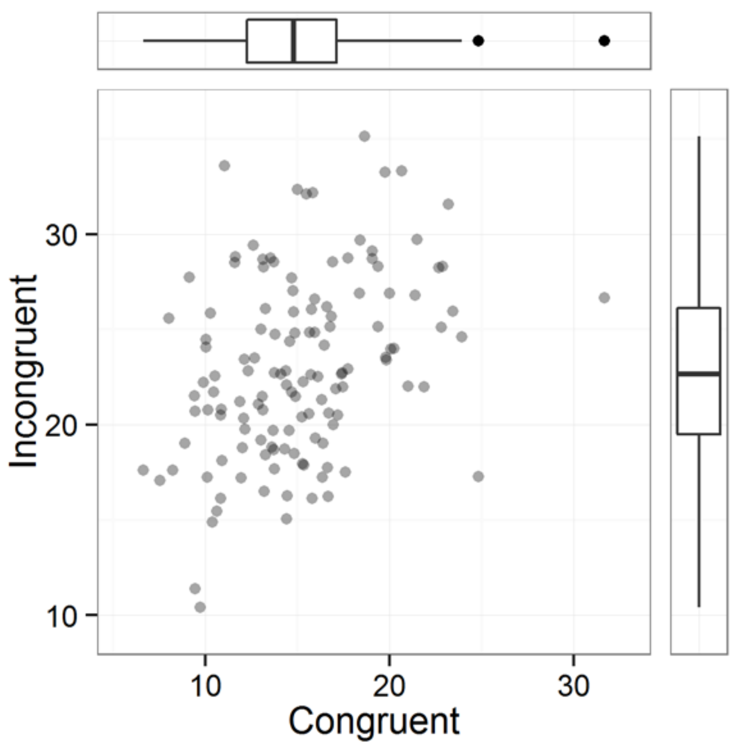
This document summarizes a comparison between two independent groups, comparing reaction times (in seconds) between the Congruent and Incongruent conditions. This script can help to facilitate the analysis of data, and the word-output might prevent copy-paste errors when transferring results to a manuscript.

Researchers can base their statistical inferences on Frequentist or robust statistics, as well as on Bayesian statistics. Effect sizes and their confidence intervals are provided, thus inviting researchers to interpret their data from multiple perspectives.

Checking for outliers, normality, equality of variances.

Outliers

Boxplots can be used to identify outliers. Boxplots give the median (thick line), and 25% of the data above and below the median (box). End of whiskers are the maximum and minimum value when excluding outliers (which are indicated by dots).



Normality assumption

The dependent t-test assumes that *difference* scores are normally distributed and that the variances of the two groups are equal. It does *not* assume the data within each measurement (so within the Congruent and Incongruent condition) are normally distributed. If the normality assumption is violated, the Type 1 error rate of the test is no longer controlled, and can substantially increase beyond the chosen significance level. Formally, a normality test based on the data is incorrect, and the normality assumption should be tested on additional (e.g., pilot) data. Nevertheless, a two-step procedure (testing the data for normality, and using alternatives for the traditional t-test if normality is violated) works well (see Rochon, Gondan, & Kieser, 2012).

Tests for normality

Yap and Sim (2011, p. 2153) recommend: "If the distribution is symmetric with low kurtosis values (i.e. symmetric short-tailed distribution), then the D'Agostino-Pearson and Shapiro-Wilkes tests have good power. For symmetric distribution with high sample kurtosis (symmetric long-tailed), the researcher can use the JB, Shapiro-Wilkes, or Anderson-Darling test." The Kolmogorov-Smirnov (K-S) test is often used, but no longer recommended, and not included here.

If a normality test rejects the assumptions that the data is normally distributed (with $p <$ .05) non-parametric or robust statistics have to be used (robust analyses are provided below).

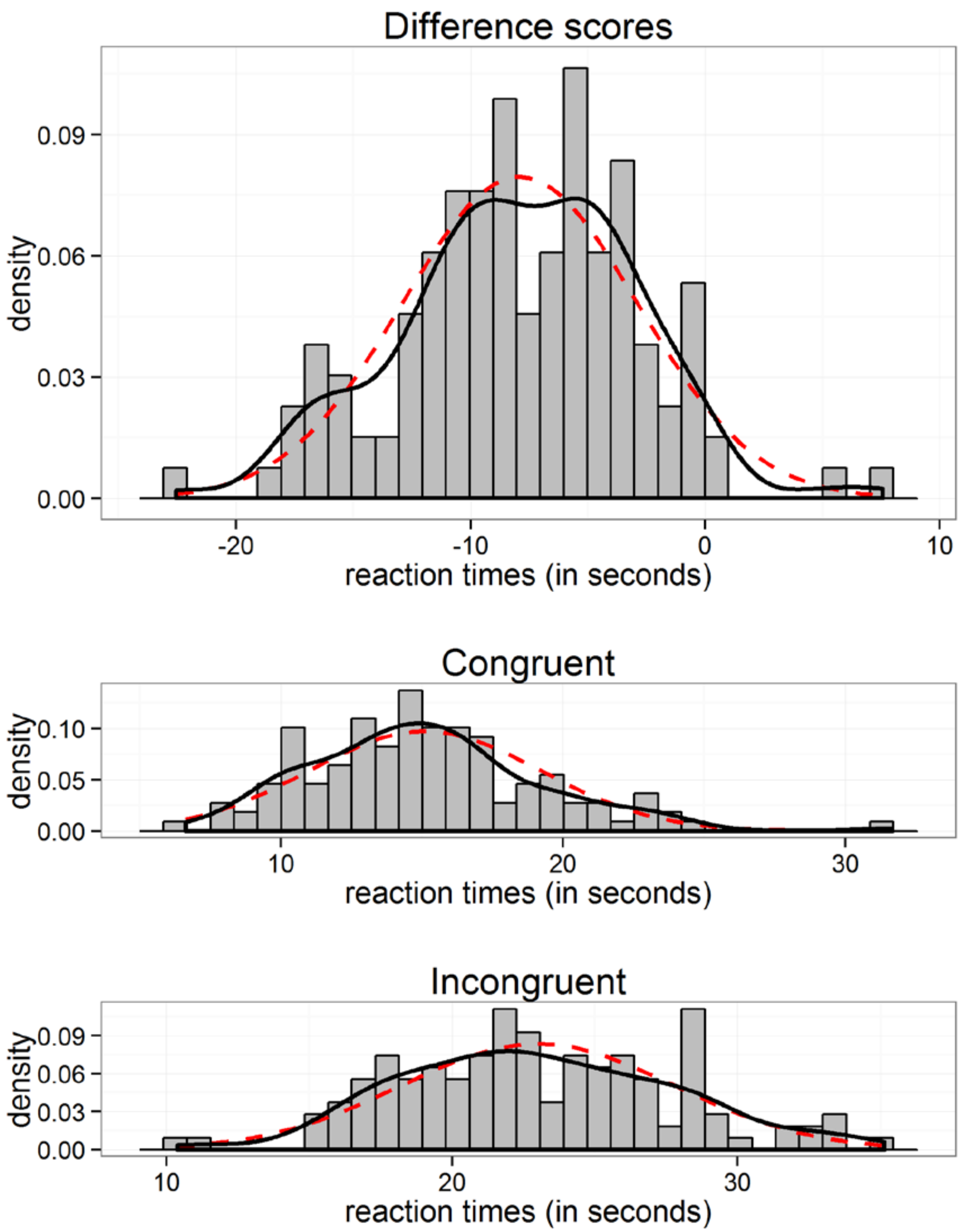
The normality assumption was rejected in 0 out of 4 normality tests (Anderson-Darling, D'Agostino-Pearson, and Shapiro-Wilk).

Test Name	p-value
Shapiro-Wilk	p = 0.442
D'Agostino-Pearson	p = 0.627
Anderson-Darling	p = 0.333
Jarque-Berra	p = 0.765

In very large samples (when the test for normality has close to 100% power) tests for normality can result in significant results even when data is normally distributed, based on minor deviations from normality. In very small samples (e.g., n = 10), deviations from normality might not be detected, but this does not mean the data is normally distributed. Always look at a plot of the data in addition to the test results.

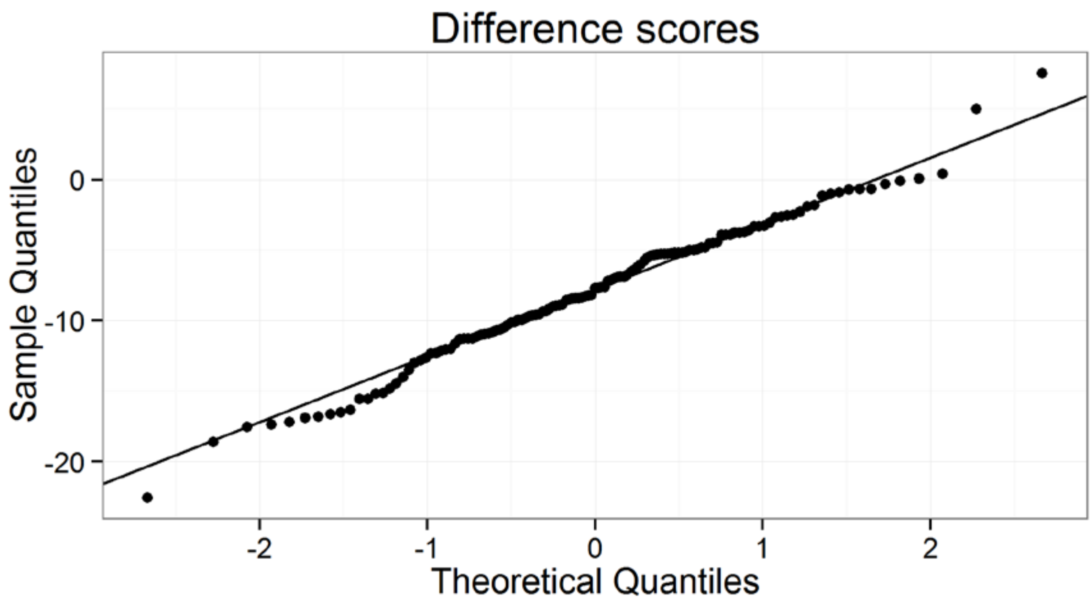
Histogram, kernel density plot (black line) and normal distribution (red line) of difference scores

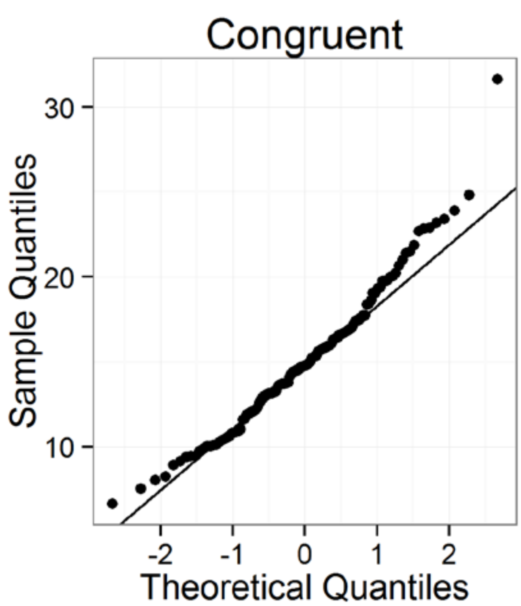
The density (or proportion of the observations) is plotted on the y-axis. The grey bars are a histogram of the difference scores. Judging whether data is normally distributed on the basis of a histogram depends too much on the number of bins (or bars) in the graph. A kernel density plot (a non-parametric technique for density estimation) provides an easier way to check the normality of the data by comparing the shape of the density plot (the black line) with a normal distribution (the red dotted line, based on the observed mean and standard deviation). For dependent t-tests, the main DV is the *difference score*, and therefore the difference score should be normally distributed.

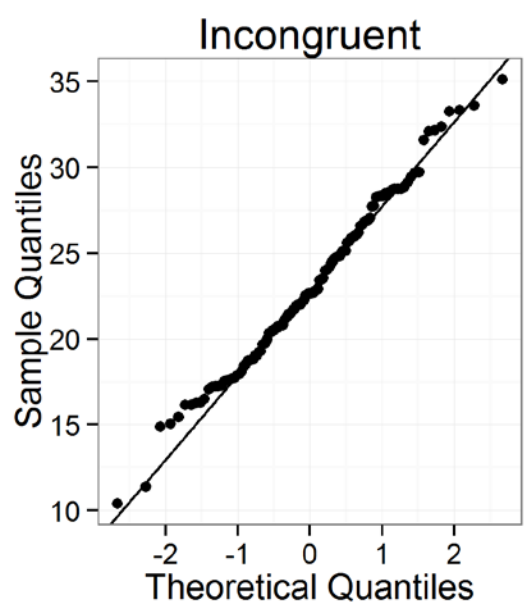


Q-Q-plot

In the Q-Q plot for the difference scores the points should fall on the line. Deviations from the line in the upper and lower quartiles indicates the tails of the distributions are thicker or thinner than in the normal distribution. An S-shaped curve with a dip in the middle indicates data is left-skewed (more values to the right of the distribution), while a bump in the middle indicates data is right-skewed (more values to the left of the distribution). For interpretation examples, see here.







Equal variances assumption

In addition to the normality assumption, a second assumption of the *t*-test is that variances in both groups are equal. The variance is the standard deviation, squared, and the assumption is thus that the variance in the Congruent condition (16.81) equals that in the Incongruent condition (22.88). Markowski & Markowski (1990) show that if sample sizes are equal, violations of the equal variance assumption do not lead to unsatisfactory performance (defined as actual significance levels falling outside a 0.03-0.07 boundary for a nominal alpha level of 0.05).

Levene's test

This equality of variances assumption is typically examined with Levene's test, although in small samples, Levene's test can have low power, and thus fail to reject the null-hypothesis that variances are equal, even when they are unequal. Levene's test for equality of variances (p = 0.038) indicates that the assumption that variances are equal is rejected (consider reporting robust statistics).

Comparing the two sets of data

Frequentist statistics

A p-value is the probability of obtaining the observed result, or a more extreme result, assuming the null-hypothesis is true. It is not the probability that the null-hypothesis or the alternative hypothesis is true (for such inferences, see Bayesian statistics below). In repeated sampling, 95% of future 95% confidence intervals can be expected to contain the true population parameters (e.g, the mean difference or the effect size). Confidence intervals are not a statement about the probability that a single confidence interval contains the true population parameter, but a statement about the probability that future confidence intervals will contain the true population parameter. Hedges' g (also referred to as $d_{unbiased}$, see Borenstein, Hedges, Higgins, & Rothstein, 2009) is provided as best estimate of Cohen's d, but the best estimate of the confidence interval is based on d_{av} (as recommended by Cumming, 2012). Hedges's g and the 95% CI around the effect size are calculated using the MBESS package by (Kelley (2007). The common language effect size expresses the probability that in any random pairing of two observations from both groups, the observation from one group is higher than the observation from the other group, see McGraw & Wong, 1992. In a dependent *t*-test, the effect size Cohen's d can be calculated by using a standardizer that controls for the correlation between observations (d_{av}) or not (d_z). Both are provided, but d_{av} (or actually it's unbiased estimate, g_{av}) is recommended. For a discussion, see Lakens, 2013. Default interpretations of the size of an effect as provided here should only be used as a last resort, and it is preferable to interpret the size of the effect in relation to other effects in the literature, or in terms of its practical significance.

Results

The mean reaction times (in seconds) (M = 15.1, SD = 4.1) of participants in the Congruent condition was smaller than the mean (M = 23, SD = 4.78) of participants in the Incongruent condition (r = 0.37). The difference between measurements is (M = -7.9, SD = 5.02), 95% CI = [-8.77;-7.04], $t(130)$ = -18.04, p < 0.001, Hedges' g = -1.76, 95% CI [-2.06;-1.48] (or d_z = -1.58, 95% CI [-1.83;-1.32]). This can be considered a large effect. The observed data is surprising under the assumption that the null-hypothesis is true. The Common Language effect size (McGraw & Wong, 1992) indicates that after controlling for individual differences, the likelihood that a persons reaction times (in seconds) in the Congruent condition is smaller than the reaction times (in seconds) in the Incongruent condition is 94%.

Bayesian statistics

Bayesian statistics can quantify the relative evidence in the data for either the alternative hypothesis or the null hypothesis. Bayesian statistics require priors to be defined. In the Bayes Factor calculation reported below, a non-informative Jeffreys prior is placed on the variance of the normal population, while a Cauchy prior is placed on the standardized effect size (for details, see Morey & Rouder, 2011). Calculations are performed using the BayesFactor package. Default interpretations of the strength of the evidence are provided but should not distract from the fact that strength of evidence is a continuous function of the Bayes Factor. A second popular Bayesian approach relies on estimation, and the mean posterior and 95% higest density intervals (HDI) are calculated following recommendations by Kruschke, (2013) based on vague priors. According to Kruschke (2010, p. 34): 'The HDI indicates which points of a distribution we believe in most strongly. The width of the HDI is another way of measuring uncertainty of beliefs. If the HDI is wide, then beliefs are uncertain. If the HDI is narrow, then beliefs are fairly certain.' To check the convergence and fit of the HDI simulations, the Brooks-Gelman-Rubin scale reduction factor for the difference score should be smaller than 1.1 (it is 1.0000213) and the effective sample size should be larger than 10000 (it is 63277). Thus, the HDI simulation is acceptable.

Results

The JZS BF_{10} (with r scale = 0.5) = 5.820268510^{33}. This indicates the data are 5.820268510^{33} (or $\log_e$ BF =77.75) times more probable under the alternative hypothesis, than under the null hypothesis. This data provides decisive evidence for H1. The posterior mean difference is -7.87, 95% HDI = [-8.73; -6.99].

Robust statistics

Values in the tails of the distribution can have a strong influence on the mean. If values in the tails differ from a normal distribution, the power of a test is reduced and the effect size estimates are biased, even under slight deviations from normality (Wilcox, 2012). One way to deal with this problem is to remove the tails in the analysis by using *trimmed means*. A recommended percentage of trimming is 20% from both tails (Wilcox, 2012), which means inferences are based on the 60% of the data in the middle of the distribution. Yuen's method can be used to compare trimmed means (when the percentage of trimming is 0%, Yuen's method reduces to Welch's t-test). The equivalent of Cohen's d for within designs is not yet available, so the explanatory effect size is reported (Wilcox & Tian, 2011).

Explanatory power (Xi, replace in the output below by the Greek lowercase Xi symbol) is the robust equivalent of omega squared (unbiased eta squared, or r squared), and thus related to d_z in size, not to d_{av}. The effect size convention of small, medium, and large corresponds approximately to Xi = 0.15, 0.35 and 0.50.

Results

Using the Yuen-Welch method for comparing 20% trimmed means showed the mean difference in reaction times (in seconds) between conditions is (M = -8.01, 95% CI [-8.04;-7.98]), which is statistically different from zero, $t(78) = -16.91$, $p < 0.001$, Xi = -1.67. The observed data is surprising under the assumption that the null-hypothesis is true. This can be considered a large effect.

Plotting data

Graph examples. In the code, you can turn different layers on and off, and change their properties by adding or removing # in front of a line of code. Displays violin plot (rotated kernal density plots) and 95% CI bars (for between designs and adjusted for within designs following Morey (2008)), individual data-points, or simple bar graphs.

***Figure 1*. Means, violin plot, and two-tiered 95% within (crossbars) and between (endpoints of lines) confidence intervals following Baguley (2012), and violin plot.**

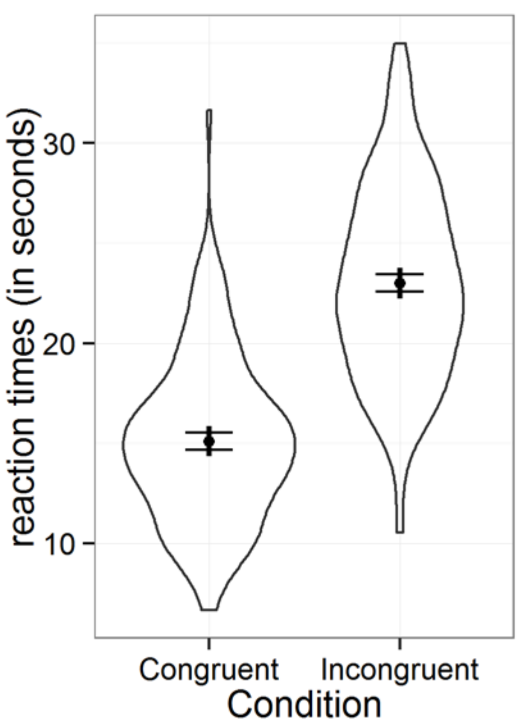


Figure 2. **Means, datapoints, and 95% CI (between & within)**

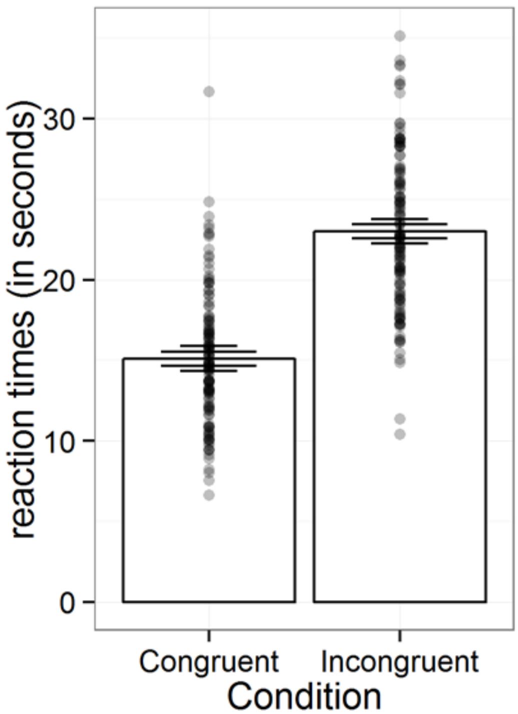
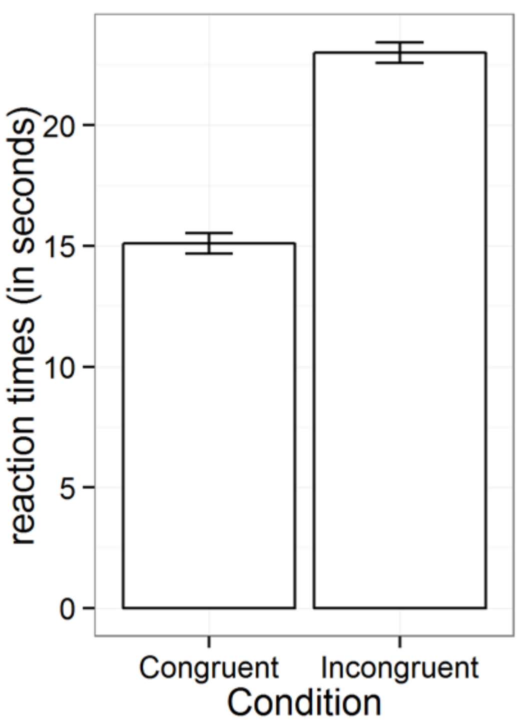


Figure 3. **Bar chart displaying means and 95% CI (between)**



References

This script uses the *reshape2* package to convert data from wide to long format, the *PoweR* package to perform the normality tests, *HLMdiag* to create the QQplots, *ggplot2* for all plots, *gtable* and *gridExtra* to combine multiple plots into one, *car* to perform Levene's test, *MBESS* to calculate effect sizes and their confidence intervals, *WRS* for the robust statistics, *BayesFactor* for the bayes factor, and *BEST* to calculate the Bayesian highest density interval.

Auguie, B. (2012). *gridExtra: functions in Grid graphics*. R package version 0.9.1, URL: http://CRAN.R-project.org/package=gridExtra.

Baguley, T. (2012). Calculating and graphing within-subject confidence intervals for ANOVA. *Behavior research methods*, *44*, 158-175.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Hoboken, NJ: Wiley.

Box, G. E. P. (1953). Non-normality and tests on variance. *Biometrika*, *40*, 318-335.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.)*. Hillsdale, NJ: Erlbaum.

Fox, J. & Weisberg, S. (2011). *An R Companion to Applied Regression, Second edition*. Sage, Thousand Oaks CA. URL: http://socserv.socsci.mcmaster.ca/jfox/Books/Companion.

Kelley, K. (2005). The effects of nonnormal distributions on confidence intervals around the standardized mean difference: Bootstrap and parametric confidence intervals. *Educational and Psychological Measurement*, *65*, 51-69.

Kelley, K. (2007). Confidence intervals for standardized effect sizes: Theory, application, and implementation. *Journal of Statistical Software*, *20*, 1-24.

Kelley, K. & Lai, K. (2012). *MBESS. R package version 3.3.3*, URL: http://CRAN.R-project.org/package=MBESS.

Kruschke, J. (2010). *Doing Bayesian data analysis: A tutorial introduction with R*. Academic Press.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t-test. *Journal of Experimental Psychology: General*, *142*, 573-603.

Kruschke, J. K., & Meredith, M. (2014). *BEST: Bayesian Estimation Supersedes the t-test*. R package version 0.2.2, URL: http://CRAN.R-project.org/package=BEST.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in psychology*, *4*.

Loy, A., & Hofmann, H. (2014). HLMdiag: A Suite of Diagnostics for Hierarchical Linear Models. R. Journal of Statistical Software, 56, pp. 1-28. URL: http://www.jstatsoft.org/v56/i05/.

McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin*, *111*, 361-365.

Micheaux, PLd. & Tran, V. (2012). PoweR. URL: http://www.biostatisticien.eu/PoweR/.

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorial in Quantitative Methods for Psychology*, *4*, 61-64.

Morey, R. D. & Rouder, J. N. (2011). Bayes Factor Approaches for Testing Interval Null Hypotheses. *Psychological Methods*, *16*, 406-419

Morey R and Rouder J (2015). *BayesFactor: Computation of Bayes Factors for Common Designs*. R package version 0.9.11-1, URL: http://CRAN.R-project.org/package=BayesFactor.

Smithson, M. (2001). Correct confidence intervals for various regression effect sizes and parameters: The importance of noncentral distributions in computing intervals. *Educational and Psychological Measurement*, *61*, 605-632.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 752-760

Wickham, H. (2007). Reshaping Data with the reshape Package. *Journal of Statistical Software*, *21*, pp. 1-20. URL: http://www.jstatsoft.org/v21/i12/.

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*. Springer New York. ISBN 978-0-387-98140-6, URL: http://had.co.nz/ggplot2/book.

Wickham, H. (2012). *gtable: Arrange grobs in tables*. R package version 0.1.2, URL: http://CRAN.R-project.org/package=gtable.

Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing*. Academic Press.

Wilcox, R. R., & Schönbrodt, F. D. (2015). *The WRS package for robust statistics in R (version 0.27.5)*. URL: https://github.com/nicebread/WRS.

Apendix A: Data & Session Information

```
alldata

##     PPNR Congruent Incongruent Year    diff
## 1      1    21.871      21.974 2013  -0.103
## 2      2    22.820      25.116 2013  -2.296
## 3      3    14.810      18.500 2013  -3.690
## 4      4    10.142      20.786 2013 -10.644
## 5      5    14.414      22.097 2013  -7.683
```

```
## 6       6    19.803       23.402 2013  -3.599
## 7       7     9.881       22.212 2013 -12.331
## 8       8    15.939       24.833 2013  -8.894
## 9       9     9.457       11.374 2013  -1.917
## 10     10     8.230       17.616 2013  -9.386
## 11     11    21.490       29.729 2013  -8.239
## 12     12    16.300       21.300 2013  -5.000
## 13     13    14.700       27.704 2013 -13.004
## 14     14    11.600       28.500 2013 -16.900
## 15     15    16.765       25.133 2013  -8.368
## 16     16    12.890       21.077 2013  -8.187
## 17     17    12.081       20.331 2013  -8.250
## 18     18    17.439       22.710 2013  -5.271
## 19     19    15.666       24.837 2013  -9.171
## 20     20    18.399       29.679 2013 -11.280
## 21     21    17.063       21.878 2013  -4.815
## 22     22    16.096       22.528 2013  -6.432
## 23     23    10.628       15.449 2013  -4.821
## 24     24    14.381       22.821 2013  -8.440
## 25     25    13.785       24.727 2013 -10.942
## 26     26    11.038       33.600 2013 -22.562
## 27     27    15.315       22.254 2013  -6.939
## 28     28    17.739       22.917 2013  -5.178
## 29     29    24.825       17.259 2013   7.566
## 30     30    20.065       23.970 2013  -3.905
## 31     31    13.137       20.761 2013  -7.624
## 32     32    13.552       28.738 2013 -15.186
## 33     33    13.121       28.694 2013 -15.573
## 34     34    12.324       22.832 2013 -10.508
## 35     35     6.636       17.598 2013 -10.962
## 36     36    15.469       32.099 2013 -16.630
## 37     37    15.967       19.289 2013  -3.322
## 38     38    14.543       19.697 2013  -5.154
## 39     39    16.332       17.231 2013  -0.899
## 40     40    15.347       17.871 2013  -2.524
## 41     41    14.984       32.352 2013 -17.368
## 42     42    15.809       16.130 2013  -0.321
## 43     43    14.106       22.661 2013  -8.555
## 44     44    19.774       23.533 2013  -3.759
## 45     45    13.200       16.500 2013  -3.300
## 46     46    13.701       28.526 2013 -14.825
## 47     47    15.274       17.935 2013  -2.661
## 48     48    14.743       27.036 2013 -12.293
## 49     49     8.032       25.590 2013 -17.558
## 50     50    11.931       17.201 2013  -5.270
## 51     51    20.664       33.328 2013 -12.664
## 52     52    13.260       18.410 2013  -5.150
## 53     53    23.417       25.966 2013  -2.549
## 54     54    13.146       28.276 2013 -15.130
## 55     55    10.031       24.467 2013 -14.436
```

```
## 56    56    16.941      19.983 2013  -3.042
## 57    57    10.843      20.492 2013  -9.649
## 58    58    11.631      28.823 2013 -17.192
## 59    59     9.128      27.731 2013 -18.603
## 60    60    10.870      20.802 2013  -9.932
## 61    61    16.601      26.202 2013  -9.601
## 62    62    12.657      23.506 2013 -10.849
## 63    63     7.525      17.075 2013  -9.550
## 64    64    23.919      24.615 2013  -0.696
## 65    65    10.375      14.871 2013  -4.496
## 66    66    13.743      17.680 2013  -3.937
## 67    67    17.405      22.651 2013  -5.246
## 68    68    10.442      21.725 2013 -11.283
## 69    69     8.900      19.034 2013 -10.134
## 70    70    13.700      18.700 2013  -5.000
## 71    71    13.671      19.681 2013  -6.010
## 72    72    12.598      29.429 2013 -16.831
## 73    73    14.908      21.464 2013  -6.556
## 74    74    19.352      25.129 2013  -5.777
## 75    75    19.359      28.297 2013  -8.938
## 76    76    19.997      26.881 2013  -6.884
## 77    77    20.237      24.007 2013  -3.770
## 78    78    16.648      16.238 2013   0.410
## 79    79    14.412      15.064 2013  -0.652
## 80    80    31.669      26.643 2013   5.026
## 81    81    14.286      18.714 2013  -4.428
## 82    82    12.110      23.431 2013 -11.321
## 83    83    13.619      18.830 2013  -5.211
## 84    84    13.741      22.715 2014  -8.974
## 85    85    14.788      25.916 2014 -11.128
## 86    86    16.819      25.677 2014  -8.858
## 87    87    11.888      21.213 2014  -9.325
## 88    88    10.516      22.556 2014 -12.040
## 89    89     9.436      20.715 2014 -11.279
## 90    90    13.256      26.072 2014 -12.816
## 91    91    18.643      35.135 2014 -16.492
## 92    92    15.827      32.161 2014 -16.334
## 93    93    13.000      25.000 2014 -12.000
## 94    94    17.600      17.513 2014   0.087
## 95    95    19.065      28.724 2014  -9.659
## 96    96    16.387      19.028 2014  -2.641
## 97    97    19.765      33.259 2014 -13.494
## 98    98    10.281      25.849 2014 -15.568
## 99    99    16.702      20.620 2014  -3.918
## 100  100    17.465      21.981 2014  -4.516
## 101  101    19.040      29.129 2014 -10.089
## 102  102    22.678      28.249 2014  -5.571
## 103  103    15.767      26.069 2014 -10.302
## 104  104    14.445      16.273 2014  -1.828
## 105  105    16.888      28.540 2014 -11.652
```

```
## 106  106    10.034      24.069 2014 -14.035
## 107  107    10.091      17.241 2014  -7.150
## 108  108    16.460      24.180 2014  -7.720
## 109  109    15.721      22.622 2014  -6.901
## 110  110    10.900      18.100 2014  -7.200
## 111  111    17.196      20.492 2014  -3.296
## 112  112    21.392      26.789 2014  -5.397
## 113  113    17.725      28.750 2014 -11.025
## 114  114    23.204      31.584 2014  -8.380
## 115  115    15.926      26.590 2014 -10.664
## 116  116    16.618      17.744 2014  -1.126
## 117  117    14.850      24.800 2014  -9.950
## 118  118    13.003      19.204 2014  -6.201
## 119  119    15.247      20.389 2014  -5.142
## 120  120     9.396      21.525 2014 -12.129
## 121  121    11.998      18.776 2014  -6.778
## 122  122    15.615      20.585 2014  -4.970
## 123  123    22.891      28.302 2014  -5.411
## 124  124    10.825      16.130 2014  -5.305
## 125  125    14.690      21.712 2014  -7.022
## 126  126    21.000      22.000 2014  -1.000
## 127  127    18.363      26.899 2014  -8.536
## 128  128     9.733      10.396 2014  -0.663
## 129  129    14.563      24.382 2014  -9.819
## 130  130    12.162      19.769 2014  -7.607
## 131  131    13.076      21.463 2014  -8.387
```

```r
sessionInfo()
```

```
## R version 3.2.0 (2015-04-16)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 8 x64 (build 9200)
##
## locale:
## [1] LC_COLLATE=Dutch_Netherlands.1252  LC_CTYPE=Dutch_Netherlands.1252
## [3] LC_MONETARY=Dutch_Netherlands.1252 LC_NUMERIC=C
## [5] LC_TIME=Dutch_Netherlands.1252
##
## attached base packages:
## [1] grid      parallel  stats     graphics  grDevices utils     datasets
## [8] methods   base
##
## other attached packages:
##  [1] BEST_0.2.2           rjags_3-15           BayesFactor_0.9.11-1
##  [4] coda_0.17-1          MASS_7.3-40          WRS_0.27.5
##  [7] MBESS_3.3.3          car_2.0-25           gridExtra_0.9.1
## [10] gtable_0.1.2         ggplot2_1.0.1        HLMdiag_0.2.5
## [13] lme4_1.1-7           Matrix_1.2-0         PoweR_1.0.4
## [16] Rcpp_0.11.6          reshape2_1.4.1
##
```

```
## loaded via a namespace (and not attached):
##  [1] formatR_1.2        nloptr_1.0.4       plyr_1.8.2
##  [4] tools_3.2.0        digest_0.6.8       evaluate_0.7
##  [7] nlme_3.1-120       lattice_0.20-31    mgcv_1.8-6
## [10] yaml_2.1.13        mvtnorm_1.0-2      SparseM_1.6
## [13] proto_0.3-10       stringr_1.0.0      knitr_1.10
## [16] MatrixModels_0.4-0 gtools_3.4.2       stats4_3.2.0
## [19] nnet_7.3-9         pbapply_1.1-1      rmarkdown_0.5.1
## [22] minqa_1.2.4        magrittr_1.5       scales_0.2.4
## [25] htmltools_0.2.6    splines_3.2.0      pbkrtest_0.4-2
## [28] colorspace_1.2-6   labeling_0.3       quantreg_5.11
## [31] stringi_0.4-1      munsell_0.4.2
```