Independent *t*-test

Daniel Lakens, D.Lakens@tue.nl

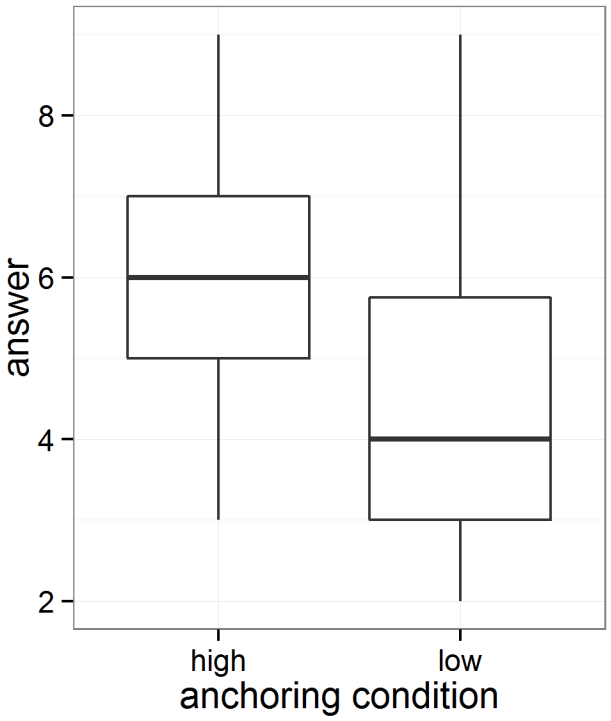
This document summarizes a comparison between two independent groups, comparing answer between the high and low conditions. This script can help to facilitate the analysis of data, and the word-output might prevent copy-paste errors when transferring results to a manuscript.

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 (whih are indicated by dots).



Normality assumption

The independent *t*-test assumes that scores in both groups (high and low) 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, seems to work well (see Rochon, Gondan, & Kieser, 2012).

Tests for normality

Four tests for normality are reported below for both groups. 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 for the high condition, and in 2 out of 4 normality tests for the low condition.

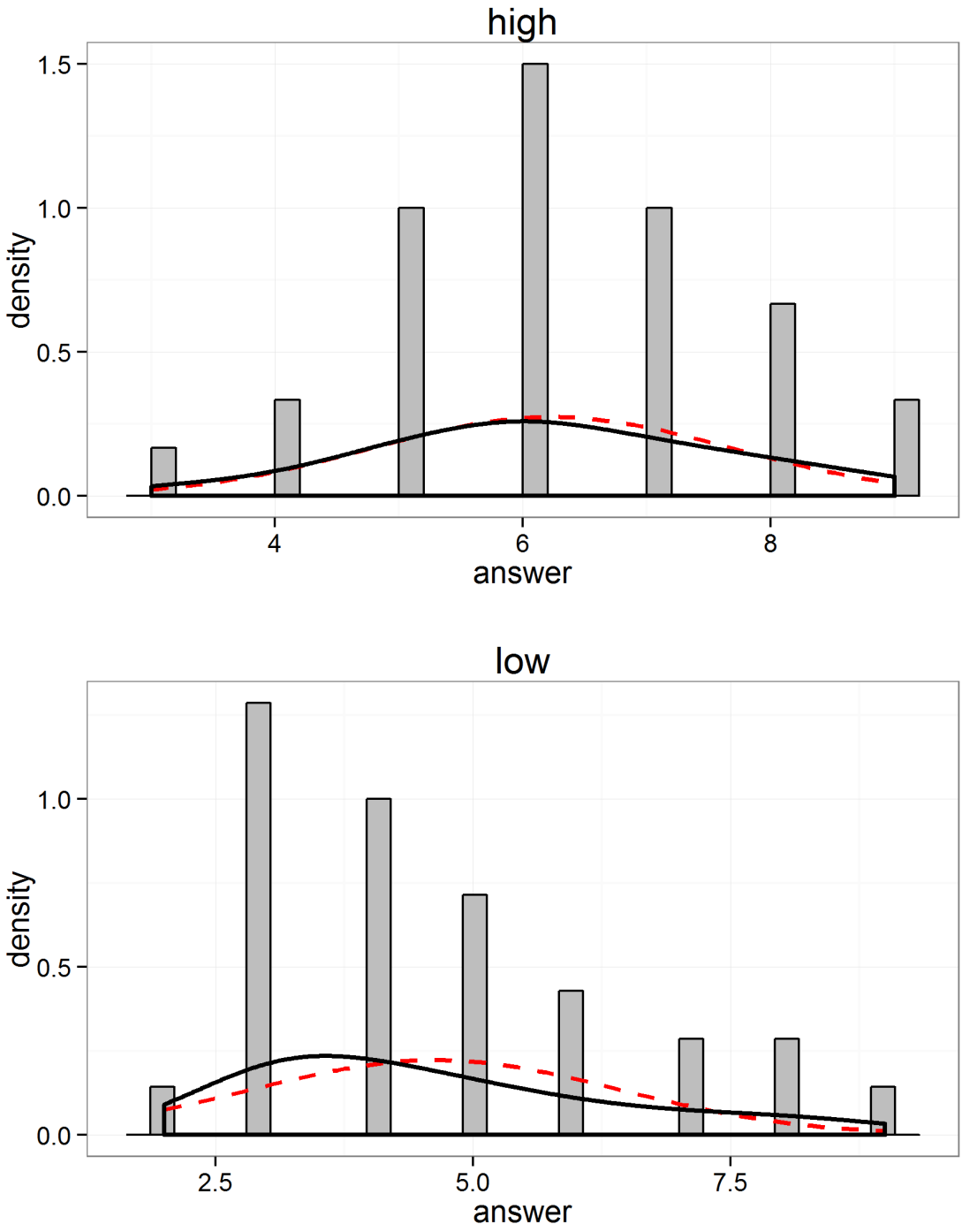
Test Name	*p*-value high	*p*-value low
Shapiro-Wilk	$p = 0.263$	$p = 0.007$
D'Agostino-Pearson	$p = 0.999$	$p = 0.133$
Anderson-Darling	$p = 0.095$	$p = 0.003$
Jarque-Berra	$p = 0.921$	$p = 0.179$

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

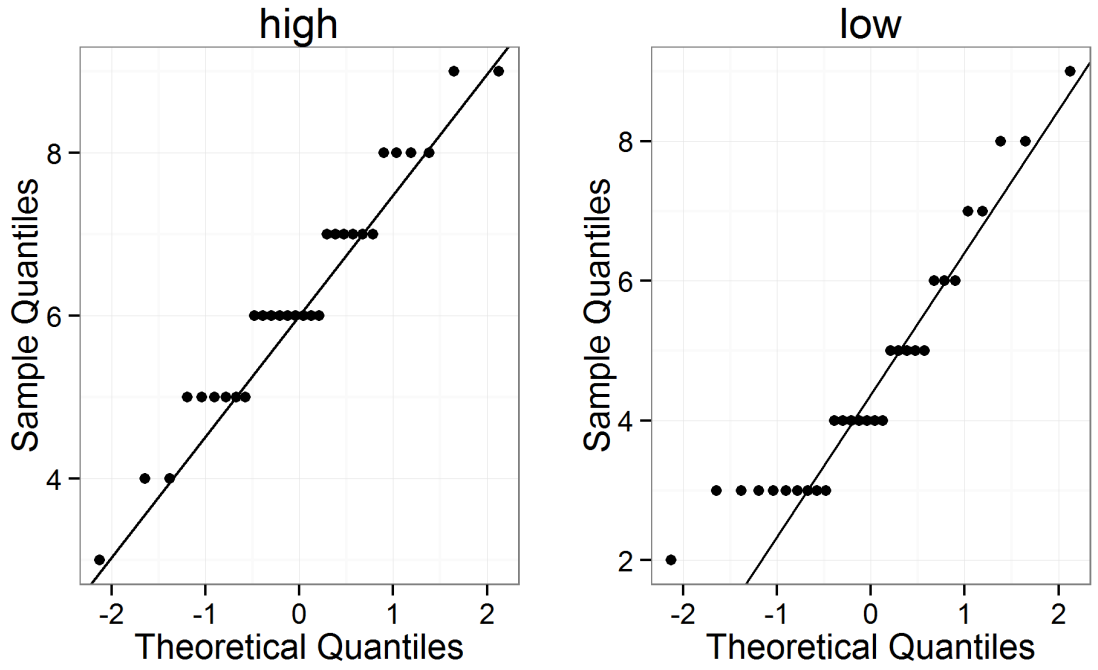
The density (or proportion of the observations) is plotted on the y-axis. The grey bars are a histogram of the scores in the two groups. Judging whether data is normally distributed on the basis of a histogram depends too much on the number of bins (or bars) in the graph. A kernel density plot (a non-parametric 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 independent *t*-tests, the dependent variables in both conditions should be normally distributed.



Q-Q-plot

In the Q-Q plots for the high and low conditions 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 Student's *t*-test is that variances in both groups are equal. As Ruxton (2006) explains: "If you want to compare the central tendency of 2 populations based on samples of unrelated data, then the unequal variance (or Welch's) *t*-test should always be used in preference to the Student's *t*-test or Mann-Whitney U test." This is preferable to the more traditional two-step approach of first testing equality o variances using Levene's test, and then deciding between Student's and Welch's *t*-test. The degrees of freedom for Welch's *t*-test is typically not a round number.

Levene's test

The equality of variances assumption is typically examined with Levene's test, although as explained above, Welch's test is used below regardless of the outcome. Levene's test for equality of variances ($p = 0.37$) indicates that the assumption that variances are equal is not rejected.

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 paramters (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 (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. 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 answer (M = 6.23, SD = 1.45, n = 30) of participants in the high condition was greater than the mean (M = 4.63, SD = 1.79, n = 30) of participants in the low condition. The difference between measurements is (M = 1.6), 95% CI = [0.76;2.44], $t(55.67) = 3.8$, $p < 0.001$, Hedges' g = 0.97, 95% CI [0.44;1.51]. 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 the likelihood that a persons answer in the high condition is greater than the answer in the low condition is 76%.

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. For a detailed explanation of an independent t-test, see this post by Richard Morey. 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 both groups should be smaller than 1.1 (For high: 1.0001434, and for low: 1.0000544) and the effective sample size should be larger than 10000 (for high: 60589, and for low: 56995). Thus, the HDI simulation is acceptable.

Results

The JZS BF_{10} (with r scale = 0.5) = 68. This indicates the data are 68 (or $\log_e$ BF =4.22) times more likely under the alternative hypothesis, than under the null hypothesis. This data provides very strong evidence for H1. The posterior mean difference is 1.66, 95% HDI = [0.79; 2.55].

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). Here, a bootstrapped version of Yuen's (1974) adaptation of Welch's two-sample test with trimmed means and windsorized variances is used that returns symmetric confidence intervals (see Keselman, Othman, Wilcox, & Fradette, 2004). Robust effect sizes and their confidence intervals are calculated using bootES by Kirby and Gerlanc (2013) following Algina, Keselman, and Penfield (2005).

Results

Using the Yuen-Welch method for comparing 20% trimmed means showed the mean difference in answer between conditions is (M = 1.89, 95% symmetric CI [0.84;2.93]), t = 4.19, p = 0.002, Robust d_t = -1.16, 95% CI = [-2.15;-0.47]). 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, individual data-points, or simple bar graphs.

Figure 1. **Means and 95% CI, and violin plot**

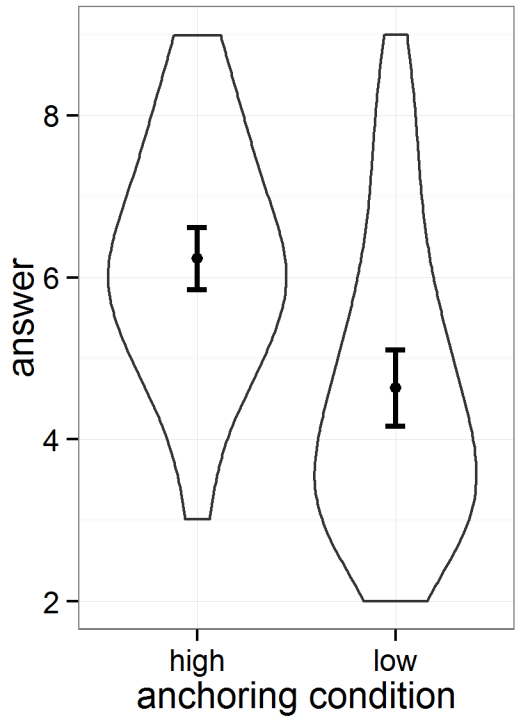


Figure 2. **Bar chart displaying means, individual datapoints, and 95% CI**

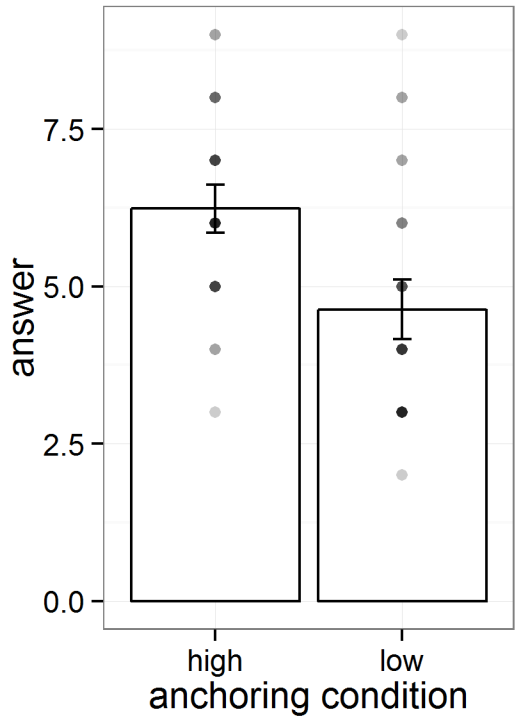
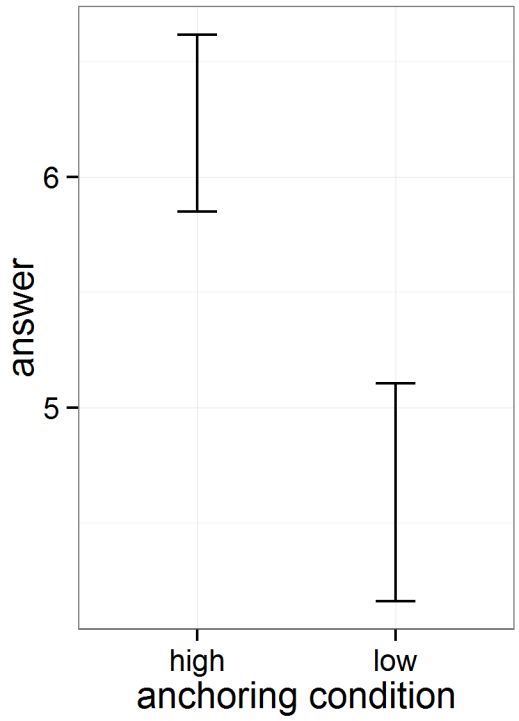


Figure 3. Bar chart displaying 95% CI



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, *bootES* for the robust effect size, *BayesFactor* for the bayes factor, and *BEST* to calculate the Bayesian highest density interval.

Algina, J., Keselman, H. J., & Penfield, R. D. (2005). An alternative to Cohen's standardized mean difference effect size: a robust parameter and confidence interval in the two independent groups case. *Psychological Methods*, *10*, 317-328.

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.

Kirby, K. N., & Gerlanc, D. (2013). BootES: An R package for bootstrap confidence intervals on effect sizes. *Behavior Research Methods*, *45*, 905-927.

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.

Rochon, J., Gondan, M., & Kieser, M. (2012). To test or not to test: Preliminary assessment of normality when comparing two independent samples. *BMC Medical Research Methodology*, 12:81.

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

Ruxton, G. D. (2006). The unequal variance t-test is an underused alternative to Student's t-test and the Mann-Whitney U test. *Behavioral Ecology*, *17*, 688-690.

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.

Yap, B. W., & Sim, C. H. (2011). Comparisons of various types of normality tests. *Journal of Statistical Computation and Simulation*, *81*, 2141-2155.

Apendix A: Data & Session Information

```
alldata
```

```
##    ppnr answer condition
## 1     1      6      high
## 2     2      7      high
## 3     3      6      high
## 4     4      9      high
## 5     5      5      high
## 6     6      6      high
## 7     7      4      high
## 8     8      8      high
## 9     9      7      high
## 10   10      6      high
## 11   11      5      high
## 12   12      8      high
## 13   13      7      high
## 14   14      5      high
```

```
## 15   15      6      high
## 16   16      6      high
## 17   17      5      high
## 18   18      8      high
## 19   19      9      high
## 20   20      6      high
## 21   21      7      high
## 22   22      5      high
## 23   23      6      high
## 24   24      4      high
## 25   25      5      high
## 26   26      6      high
## 27   27      3      high
## 28   28      7      high
## 29   29      8      high
## 30   30      7      high
## 31   31      3       low
## 32   32      4       low
## 33   33      3       low
## 34   34      3       low
## 35   35      3       low
## 36   36      4       low
## 37   37      3       low
## 38   38      4       low
## 39   39      4       low
## 40   40      4       low
## 41   41      5       low
## 42   42      5       low
## 43   43      3       low
## 44   44      5       low
## 45   45      4       low
## 46   46      3       low
## 47   47      6       low
## 48   48      3       low
## 49   49      2       low
## 50   50      6       low
## 51   51      6       low
## 52   52      7       low
## 53   53      8       low
## 54   54      8       low
## 55   55      7       low
## 56   56      5       low
## 57   57      5       low
## 58   58      9       low
## 59   59      4       low
## 60   60      3       low
```

```
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] gridExtra_0.9.1      gtable_0.1.2         BEST_0.2.2
##  [4] rjags_3-15           BayesFactor_0.9.11-1 coda_0.17-1
##  [7] bootES_1.01          boot_1.3-16          WRS_0.27.5
## [10] MBESS_3.3.3          car_2.0-25           HLMdiag_0.2.5
## [13] lme4_1.1-7           Matrix_1.2-0         PoweR_1.0.4
## [16] Rcpp_0.11.6          ggplot2_1.0.1
##
## 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         reshape2_1.4.1      magrittr_1.5
## [25] scales_0.2.4        htmltools_0.2.6     MASS_7.3-40
## [28] splines_3.2.0       pbkrtest_0.4-2      colorspace_1.2-6
## [31] labeling_0.3        quantreg_5.11       stringi_0.4-1
## [34] munsell_0.4.2
```