Understanding and reporting output of the SCED package

Ian Hussey

*Version 1.1 – 2018/11/14*

*Author note:* Department of Experimental-Clinical and Health Psychology,Ghent University, Belgium; [ian.hussey@ugent.be](mailto:ian.hussey@ugent.be)

**Abstract**

The SCED package for R provides a simple workflow for the analysis of A-B design (i.e., pre-post) Single Case Experimental Design research (Hussey, 2018). Although written in R, the SCED package is designed to be accessible even for R novices: the four key components of a SCED analysis (analyzing and plotting at the subject level, meta analyzing across participants, and printing a forest plot of meta analyzed effect sizes) can each be conducted with a single line of code. This document summarizes the statistics and plots outputted by these functions, and is intended to facilitate writing up results obtained using this package. A citable template for reporting SCED results in a manuscript is also provided.

**Description of analytic strategy**

**Preregister your decision making strategy**

The SCED package returns multiple metrics for hypothesis testing (e.g., *p* values, CIs on effect sizes) at the participant and also at the group level (via meta analysis of effect sizes Ruscio’s A or Hedges’ *g*). Given that multiple metrics are returned, researchers should preregister which participant level and meta analysis metrics they will use for decision making. On the basis of a power analysis simulation study that I have conducted but not yet reported, the most robust and powerful metrics at the participant level are either permuted *p* values or the confidence intervals on Ruscio’s A. At the group level, meta analyzed Ruscio’s A is likely to be more robust than Hedges’ *g*.

**Hypothesis tests via permutation tests**

A traditional within-sample t test takes the multiple data points from each condition and reduces them to a set of values that summarize this collection of data points. This is referred to as parameterization, e.g., where a dozen data point are summarized as a mean and SD. This parameterization requires making a number of assumptions that may not be the case, e.g., that the data points that are used to create a mean are normally distributed, or that SDs are equivalent between groups. Historically, tests that rely on parameterized data (i.e., parametric tests) were developed because they provided a useful mathematical shortcut when these tests would be worked out by hand or with very limited computing power.

Given modern computing power, these mathematical shortcuts are no longer necessary for may forms of analysis. Permutation tests are non parametric and are instead calculated via a brute force resampling method. Loosely speaking, if inferential statistics were being developed from scratch and one wanted the answer that a p value provides (i.e.. what is the probability of observing data at least as extreme as that observed if the null hypothesis is true), but this time you had modern computing power at your fingertips, you would have created permutation tests and never needed to make the watered down versions that are parametric t tests.

Permutation tests are a form of exact test or resampling test (related to bootstrapping) where data labels are exchanged multiple times. For example, imagine you have data points from 1 to 10 belonging to conditions A and B in the order AAAAABBBBB. Rather than assume what a null effect looks like via parameterization, a permutation test will calculate how extreme your data is in terms of how else the conditions could be assigned. E.g., it will re-label the data as BBBBBAAAAA, ABABABABAB, BBAABBAABA, and thousands of other combinations, pool these combinations together, and observe the percentile in which your real data lies in in terms of its extremity. As such, this provides an exact test of the probability of observing such data (i.e., a *p* value).

Permuted p values are particularly useful for SCED research because they contain no assumptions about the data, where SCED data frequently violates such parametric assumptions. For further reading, the [Wikipedia page on resampling statistics](https://en.wikipedia.org/wiki/Resampling_(statistics)) is a good place to start.

**Robust effect sizes**

In order to quantify the magnitude of any change, three robust effect sizes are calculated. First, the median difference between conditions. Medians are robust relative to means, have simple interpretation, and do not suffer from a ceiling effect (maximum value).

Second, Hedges' *g* values are reported for the sake of reader familiarity (Hedges, 1981). These are a standardized difference score similar to Cohen's *d* (Cohen, 1988) that includes a bias adjustment for small sample sizes, which frequently applies to SCED data. They have the same cut-off scores for interpretation. In order to make them more robust, the SCED package reports bootstrapped median Hedges' *g* effects sizes plus their bootstrapped 95% confidence intervals. This use of bootstrapping via case removal reduces the influence of outliers, mitigating violations of Hedges' *g*'s parametric assumptions (e.g., normality of data in each condition, equivalent SDs, and equal number of data points in both conditions, all of which are routinely violated in SCED data). It should be emphasized that *g* is reported for familiarity only and not recommended as the primary way to interpret the effect size. Aside from violations of its assumptions, its interpretation is also not actually that clear: technically, it is the bootstrapped, bias-corrected, difference between conditions as a proportion pooled deviation in those conditions. This is not that useful to a clinician or policy maker.

Third, Ruscio's A values are calculated (Ruscio, 2008). Ruscio's A is not that popular as a standardized effect size, but it probably should be. One trivial and unfortunate reason for this is that although it has existing in slightly different forms for years it has gone by different names, masking its popularity between domains (e.g., the Common Language Effect Size, the Probability of Superiority, the Area Under the Receiver Operating Characteristic Curve, a special case of the Probabilistic Index Model, among others). Ruscio's A is non-parametric and treats the data as ordinal rather than continuous. Its definition, and indeed its calculation via permutation, is "the probability that a randomly chosen data point in condition B is larger than a randomly chosen data point in condition A"; or, more loosely, the probability that the intervention produces improvement. Due to a combination of its very high robustness and its ease of interpretation even for non-experts, Ruscio's A is an excellent standardized effect size. Its one drawback is that it suffers from a ceiling effect: if all data points in time point B are higher than time point A (i.e., probability = 1.0), it is not possible to distinguish between a very large effect size and a extremely large one. This is overcome by also reporting the median difference. E.g., one can then conclude that the probability of superiority (Ruscio's A) is 1.0 (median difference = 4.5). For additional robustness, the SCED package reports the median Ruscio's A as well as the 95% confidence intervals on this estimate, both bootstrapped via case removal and using the percentile method for confidence intervals. Notionally, the confidence intervals on Ruscio's A could be employed for decision making purposes rather than permuted p values. This should be decided before data collection (e.g., in your study's preregistration) in order to limit researchers' degrees of freedom, as the two will not agree in all cases when the number of data points is very low in one or both conditions.

**Meta analysis**

The above provide hypothesis test and effect sizes for individual participants in a SCED study. In order to pool results across participants, the SCED package also allows for Ruscio's A and Hedges’ *g* effect sizes to be subjected to a Random Effects Meta Analysis, following best practices. Although Ruscio’s A is fully non-parametric, for ease of meta analysis we assume that the underlying effect varies normally between participants (i.e., the meta analysis model employs a Gaussian link function). This provides both 95% confidence intervals (i.e., estimates of the true population effect size) and 95% credibility intervals (i.e., the range of effect sizes likely to be observed in future participants). Finally, this also provides information about the heterogeneity observed between participants (i.e., estimates of Q, I2, and H2). It is also useful to provide an unstandardized effect size in order to get a sense of the real world change due to the intervention. This is particularly the case given Ruscio’s A potential for ceiling effects. In order to select the maximally robust unstandardized effect size, the SCED package returns the median median-difference. This refers to the median value between participants of the median value between A and B phases within participants. Put another way, the median participant demonstrated this median change due to the intervention.

**Plotting**

A plot of SCED data is produced for each participant. This includes their raw data points, a dashed vertical line to indicate when the intervention was performed, dashed horizontal lines to indicate the median value for each condition, and linear regression lines fitted to both conditions. The latter can be useful to visually diagnose trends at baseline or follow-up that may be important to qualifying the results. E.g., if there is a trend towards improvement at baseline then differences between the conditions may not be due to the intervention. If improvement at baseline may be due to method factors (e.g., if repeated presentation of some items every day acts as a mini intervention itself) then this can be mitigated by staggering the intervention time between participants. There are quantitative methods to diagnose trends within time points (e.g., magnitude of linear regression slope, significance of linear regression slope, significance of a non parametric linear regression slope such as the Theil-Sein slope). However, all of these suffer from multiple issues including violated parametric assumptions, arbitrary magnitude cutoffs, multiple testing corrections, and paradoxical power implications (e.g., where more data increases power, punishing the researcher for collecting additional data collection to find a stable baseline). As such, the issue of baseline trends has no robust quantitative solution to date, to my knowledge. In its absence, visual trends or slope values can be used. Standardized linear regression (standard beta) coefficients are reported for both the A and B experimental phases.

Additionally, a forest plot of the meta analyzed standardized effect size can be produced. This is a standard way to present meta analysis results, and provides both 95% confidence intervals and 95% credibility.

**Example of how to conduct a SCED analysis in R**

See the demo.Rmd R markdown file in the vignettes folder.

**Example of how to report SCED results**

*Below is a suggestion for how to present results from the SCED package in a manuscript. A copy of the results table and SCED data plot should be included. A copy of the meta analysis forest plot can optionally be included – these are less typical but are informative. Please cite the SCED package if you use or adapt this (see references section).*

The R package SCED was used to analyse and plot the data (Hussey, 2018) in conjunction with the metafor package (Viechtbauer, 2010). For each participant, *p* values were calculated via robust, non-parametric permutation tests. Three robust effect sizes were also calculated: 1) median difference between conditions, 2) Ruscio's A (Ruscio, 2008) – somewhat confusingly this is also referred to with may other names such as the Common Language Effect Size (McGraw & Wong, 1992), the Probability of Superiority (Ruscio & Mullen, 2012), Nonoverlap of All Pairs (Parker & Vannest, 2009), or as an instance of Probabilistic Index Modeling (Thas, De Neve, Clement, & Ottoy, 2012), and 3) Hedges' *g*. The latter, a version of Cohen’s *d* corrected for use with a small number of data points, is calculated for the sake of reader familiarity but is acknowledged to have parametric assumptions that are routinely violated by SCED data. Ruscio's A is a fully non-parametric effect size with very simply interpretation: it is “the probability that a randomly chosen data point in condition B is larger than a randomly chosen data point in condition A”; or, somewhat loosely, the probability of improvement for that participant. Both Hedge's *g* and Ruscio's A were calculated via robust estimation methods: we report the median bootstrapped value via case removal along with its 95% confidence intervals via the bias corrected and accelerated (BCA) method.

For each participant, trends at baseline were diagnosed via [visual inspection of the plotted data (see Figure XXX)/the calculation of standardized beta linear regression coefficients with a cutoff value of 0.XX]. [No] evidence of trends at baseline was observed. [*Where trends are visible and differ between participants, two meta analyses could be conducted, with and without these participants.*] Visual inspection of the SCED data also indicated [clear evidence of improvement in scores after intervention in X of Y participants].

As illustrated in Table XX, statistically significant improvement was found in X of Y participants. Standardized effect sizes were then meta analyzed across participants. Probability values (i.e., Ruscio's A) were converted to log-odds ratios and subjected to a random effects meta analysis. Meta analytic p value, estimate of the standardized effect size, its confidence intervals, and its credibility intervals were calculated. Whereas confidence intervals (CI) refer to the estimate of the true value of Ruscio's A across participants (i.e., estimate the point effect size), credibility intervals (CR) refer to estimates of the values of Ruscio's A that are likely to be observed across participants in similar future studies. Results a meta analytic standardized effect size of Ruscio's A = 0.755, 95% CI [0.642, 0.842], 95% CR [0.537, 0.892] and an unstandardized robust effect size of median median-difference XX. This refers to the median value between participants of the median value between A and B phases within participants. Put another way, the median participant demonstrated this median change due to the intervention. Finally, meta analysis demonstrated [no] evidence of heterogeneity between participants, *Q*(df = 4) = 6.99, *p* = 0.14, tau2 = 0.17, I2 = 47.36, H2 = 1.90 (where *p* < 0.05 indicates heterogeneity). This suggests that participants responded to the intervention in a comparable manner and that results can be appropriately generalized across participants.

References

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

Hedges, L. V. (1981). Distribution Theory for Glass’s Estimator of Effect size and Related Estimators. *Journal of Educational Statistics*, *6*(2), 107–128. https://doi.org/10.3102/10769986006002107

Hussey, I. (2018). *SCED: An R package for the robust analysis, visualization, and meta analysis of A-B Single-Case Experimental Design data*. Retrieved from https://github.com/ianhussey/SCED

McGraw, K. O., & Wong, S. (1992). A common language effect size statistic. *Psychological Bulletin*, *111*(2), 361.

Parker, R. I., & Vannest, K. (2009). An Improved Effect Size for Single-Case Research: Nonoverlap of All Pairs. *Behavior Therapy*, *40*(4), 357–367. https://doi.org/10.1016/j.beth.2008.10.006

Ruscio, J. (2008). A probability-based measure of effect size: Robustness to base rates and other factors. *Psychological Methods*, *13*(1), 19–30. https://doi.org/10.1037/1082-989X.13.1.19

Ruscio, J., & Mullen, T. (2012). Confidence intervals for the probability of superiority effect size measure and the area under a receiver operating characteristic curve. *Multivariate Behavioral Research*, *47*(2), 201–223.

Thas, O., De Neve, J., Clement, L., & Ottoy, J.-P. (2012). Probabilistic index models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *74*(4), 623–671. https://doi.org/10.1111/j.1467-9868.2011.01020.x

Viechtbauer, W. (2010). Conducting Meta-Analyses in R with the metafor Package. *Journal of Statistical Software*, *36*(3). https://doi.org/10.18637/jss.v036.i03