{SCED}: An R package for simple and robust visualization, analysis, and meta-analysis of A-B Single Case Experimental Designs

Ian Hussey

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.

Single Case Experimental Designs (SCED) treat individual participants as individual experiments, collect data from participants at multiple time points, and emphasise fine grain experimental control over individuals' behaviour (REFs). SCEDs therefore stand in comparison to the current dominant paradigm of groups deign studies and Randomized Controlled Trials for creating knowledge and informing clinical decision-making. In a recent review, Duan, Kravitz and Schmid (2013, p.21; see also Deaton & Cartwright, 2018) summarized three important critiques of RCTs that are overcome by SCEDs: "(1) it is practically impossible to conduct standard parallel-group randomized controlled trials (RCTs) to address all clinically important questions, even those restricted to comparative effectiveness of drugs and devices; (2) clinical evidence generated in those RCTs has poor generalizability and therefore limited applicability to real patients seen in ordinary practices; and (3) treatments shown to be safe and effective on average may deliver an uneven mix of risks and benefits to individual patients, a problem known as heterogeneity of treatment effects (HTE)." For these reasons, SCED research has seen a resurgence of interest in recent years (REFs).

One apparent barrier to the adoption of SCED is the availability and ease of use of analytic workflows. In contrast with SCED training and resources, even undergraduate training in social and medical sciences often provides introductions to simple within and between groups factorial designs and analysis strategies such as *t*-tests and ANOVAs. This has arguably led to a normalization of RCT designs relative to SCED alternatives, despite the potential benefits of the latter. Recent articles have therefore emphasized the need to make SCED analytic workflows accessible, and to evaluate SCED approaches not only on their statistical appropriateness but also the ease with which researchers can access and use such methods (Manolov & Moeyaert, 2017). While such articles have clear practical value, they highlight that a comprehensive analysis of SCED data (e.g., visualization, single subject quantitative analysis, and meta-analysis of effect sizes) can often multiple different software programs and packages (e.g., Excel, R

and SPSS). Existing efforts to integrate and simplify workflows are encouraging (e.g., Bulté & Onghena, 2013) but can likely be further improved, or target different use cases than the SCED package here.

Design philosophy and intended use cases

The SCED package was created with two main use cases in mind. First, to lower the barriers to entry to SCED for experimentalists or trialists who may have until now employed between groups experiments or Randomized Controlled Trials. One rationale for doing so is to lower research costs by decreasing the number of individuals required to provide high quality evidence for a manipulation or intervention, maximizing the return on investment for finite research funds. In doing so, accessible use of SCED is hoped to orientate researchers towards the fine grain experimental control of individual's behaviour. Second, to enable experimentalists or trialists to conduct research where it was difficult to conduct well powered research using traditional between groups and RCT designs. For example, when studying difficult to access populations, when operating with limited funding and administrative resources, or in contexts where such research has typically been more difficult to conduct (e.g., individual clinics or even practitioners who wish to create an evidence base for their practice).

The SCED package was developed according to the following philosophy and indented user base in mind. Specifically, that it would:

1. Be free and Open Source code.
2. Make use of other popular, well-validated Open Source packages. This was for two complementary reasons. First, to maximize the generalizability of users' skills between different areas. For example, data structures are Tidy Data compliant (Wickham, 2014) and therefore and are interacted with in a similar way to other popular R packages for data analysis, such as for the simple application of ANOVAs (Lawrence, 2016) or mixed effects models (Bates, Mächler, Bolker, & Walker, 2015). Second, to leverage popular and well tested packages in order to outsource good design and implementation choices to experts in those domains. For example, graphing via ggplot2 and the grammar of graphics (Wickham, 2010, 2016), efficient bootstrapping of robust effect size confidence intervals by statisticians in that area (Ruscio, 2008; Ruscio & Mullen, 2012).
3. Specifically cater to A-B SCED designs. Although a wide variety of other SCED designs are possible and indeed encouraged (Tate, Perdices, Rosenkoetter, Shadish, et al., 2016), relatively simple pre-post SCED designs are the most likely and accessible designs for the intended use cases.
4. Consciously limit the range of experimenter degrees of freedom available to researchers in the analysis of their data. Choices among which metrics and tests are included in the package were selected on the basis of 1) ease of interpretation of the results, 2) robustness to violation of parametric assumption, which are routinely violated in SCED data, 3) clarity of analytic workflow (e.g., clear recommendations for what tests should be run in what order, and how the results of one test impact others). Importantly, no attempt was made to be exhaustive in the implementation of possible methods that would satisfy all users. Users who have other specific interests, for example, in other plotting methods or effect size metrics, are expected to be either be a) more advanced users who can adapt the R code for their needs, or b) less experienced used who are looking to employ a given method "blindly" based on it being common in the literature, but agnostic to whether it is justifiably superior to the methods included in the package. For example, the Percent Non-overlapping Difference (PND) is a common SCED effect size but is less robust the Ruscio's A effect size

included in the package (REF). This is discussed further below. Of course, this is not to say that the methods implemented in the SCED package are final and objectively best methods for all SCED studies, only that they meet the intended use cases outlined above more closely than alternatives. Other methods could be added in future based on suggestions from the community – or indeed written by others and added to the package independently, given that the code is open source.

5. Be easy to use for non-experts. Following the above point, the workflow between specific analyses and tests should be clear to non-expert users. Implementation of the workflow as a whole likewise be easy for non-expert users. Currently, the workflow is implemented as a single RMarkdown script (SCED.Rmd), which only requires the user to 1) properly format their data, 2) download and install the freely available RStudio program, and 3) run SCED.Rmd inside RStudio by clicking the "knit" button. All output is created in a html file that can be viewed in any web browser, from which results can be pasted into manuscripts for publication. In the future, I hope to create a web app that will allow users to simply upload their data have results returned to them in one click, by passing the need to install any software.

Experimental designs

This package deals with the *analysis* of A-B SCED data. In designing, collecting and communicating the results of SCED data, readers are encouraged to consider the SCRIBE reporting guidelines (Tate, Perdices, Rosenkoetter, McDonald, et al., 2016; Tate, Perdices, Rosenkoetter, Shadish, et al., 2016), which are the equivalent of the CONSORT guidelines for RCTs. In order that the study be a true SCED rather than merely a 'single case methodology' researchers should strongly consider employing a multiple baseline design where the introduction of the intervention is staggered across time points between participants (Tate, Perdices, Rosenkoetter, Shadish, et al., 2016). Readers are also encouraged to read the What Works Clearinghouse guidelines for SCED research (2010), which contain many valuable procedural recommendations such as the minimum number of participants and measurement time points.

I have conducted but not yet written up a simulation study that demonstrates the statistical power of the hypothesis testing strategies employed in the SCED package, and under a range of plausible experimental designs (e.g., true effect size, number of time points both before and after intervention). These suggest that the methods included in the package are superior to other common and recent methods (e.g., autoregressive Bayes Factors) as well as informing the recommendations between the metrics included in the package (e.g., permuted p values and bootstrapped CIs on Ruscio's A are shown to have greater power than CIs on Hedge's g).

Analytic strategy

Quantitative analysis methods

Some authors argue that when sufficient experimental control is exerted over participants' behaviour quantitative analysis is redundant (REF). However, the absence of quantitative analysis continues to be a barrier to the acceptability of SCED evidence. Furthermore, research has shown visual inspection to have low inter-rater reliability, and to be poor at detecting potentially important properties of data such as autocorrelation (Ottenbacher, 1990; Park, Marascuilo, & Gaylord-Ross, 1990; Ximenes, Manolov, Solanas, & Quera, 2009). As such, if we are to exert stimulus control over not only our participants' behaviour but also our colleagues', then it is pragmatically useful to embrace quantitative methods.

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 strongly consider preregistering which participant level and meta-analysis metrics they will use for decision making and inferences, and which others will be reported but not used for decision making. This can be done easily on the Open Science Framework or other such services.

On the basis of a power analysis simulation study that I have conducted but not yet published, 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.

Robust hypothesis testing. 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 makes a number of assumptions that may not be the case, e.g., that the data points are normally distributed and therefore adequately summarized means and standard deviations, or that standard deviations are equivalent between conditions. Historically, tests that rely on 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 arguably no longer necessary for may types of analysis. Permutation tests represent a high quality alternative: these tests are a) fully non parametric and b) do not compare the observed distribution with an unobserved null distribution. Instead, permutation tests are calculated using 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 might have started with permutation tests in the first place.

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 compare the parameterizations of this distribution against an unobserved null distribution (as in a traditional within samples t test), a permutation test instead calculates how extreme your data is in terms of the actual condition assignment compared to many other potential condition assignments. E.g., it will re-label the same data points as belonging to different conditions, such as BBBBBAAAAA, ABABABABAB, BBAABBAABA, and thousands of other combinations. It will then 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 this data (i.e., a p value) with these condition assignments compared to others.

Permuted p values are particularly useful for SCED research because they contain no assumptions about the distribution of the data, given that SCED data frequently violates such parametric assumptions. See Nichols and Holmes (2002) for an accessible introduction. Various forms of permutation tests have been recommended for quantitative analysis of SCED data for over 25 years (e.g., Onghena & Edgington, 1994). The method used by the SCED package uses the popular R package coin. Specifically, all data points are considered mutually interchangeable (i.e., unordered) when generating alternative assignments. While this point about the appropriateness of treating all data points as exchangeable (vs. using only assignments that were experimentally plausible, i.e., a "randomisation test") is a matter of debate (e.g., Bulté & Onghena, 2008), it is defendable on the grounds that the most commonly used effect sizes in SCED research (e.g., Hedges' g,

Percentage of Non-overlapping Difference, Percentage Exceeding the Median, etc; see Parker, Vannest, & Davis, 2011) make the same assumption. Thus, the hypothesis testing and effect size estimation methods employed in the SCED package make congruent assumptions.

Robust effect size metrics. In order to quantify the magnitude of any change, three robust effect sizes are calculated: medians. bootstrapped Hedges' *g*, and Ruscio's A. First, the median difference between conditions. Medians are robust relative to means, have simple interpretation, and do not suffer from a ceiling effect (i.e., maximum value).

Second, Hedges' *g* values are reported for the sake of reader familiarity (Hedges, 1981). These are a standardized difference score very similar to Cohen's *d* (Cohen, 1988) that includes a bias adjustment for small sample sizes, which typically applies with SCED data. They have the same cut-off scores for interpretation (e.g., small $\geq$ 0.2, medium $\geq$ 0.5, large $\geq$ 0.8, very large $\geq$ 1.20, huge $\geq$ 2.0: Cohen, 1988; Sawilowsky, 2009). While the cutoff values for Hedges' g will be familiar to many readers, this effect size is relatively uninformative with regards to the real world size of the effect. For example, men are 5 inches taller than women on average (unstandardized effect size), but this has little intuitive correspondence with its standardized form (Cohen's d = 1.72: Ridgway, 2013). Furthermore, both *d* and *g* make the same parametric assumptions discussed above, which are routinely violated in SCED data. In order to increase robustness, bootstrapped median Hedges' *g* plus its bootstrapped 95% confidence intervals are calculated (via case removal BCA method, using the bootES package: Kirby & Gerlanc, 2013). This reduces the influence of outliers, mitigating violations of parametric assumptions. However, it should be emphasized that Hedges' *g* is calculated primarily for the sake of reader/reviewer familiarity, but is not the recommended effect size metric. Aside from violations of its assumptions, its interpretation is also not actually that clear: technically, it is the bootstrapped, bias-corrected difference between conditions' means as a proportion pooled deviation in those conditions. This is usually not that useful to a clinician or policy maker.

In order to provide a standardized effect size metric that is both robust and interpretable, the SCED package also calculates Ruscio's A values (Ruscio, 2008). Ruscio's A is currently not that common as a standardized effect size, but it probably should be. One trivial and unfortunate reason for its lack of popularity is that although it has been referred to using different names by different authors, masking its actual popularity. For example, it and its slight variants have been called the Common Language Effect Size (McGraw & Wong, 1992), the Probability of Superiority (Ruscio & Mullen, 2012), the Area Under the Receiver Operating Characteristic Curve (when the DV is binary: Egan, 1975), the Probabilistic Index (Acion, Peterson, Temple, & Arndt, 2006; Thas, De Neve, Clement, & Ottoy, 2012), Non-overlap All Pairs (Parker & Vannest, 2009), the Dominance Statistic, Mann-Whitney's U, and others. Researchers should feel free to use whatever label for this statistic they see fit: I am in the habit of referring to it as Ruscio's A, but the Probability of Superiority is also a good descriptive label. Annotation in results sections could be A for Ruscio's A or P(A<B) for the probability of superiority (of B over A).

Ruscio's A is fully non-parametric and treats the DV 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". More loosely, this is the probability that an organism is likely to produce better scores after an intervention than before. Due to a combination of its high robustness and its ease of interpretation even for non-experts, Ruscio's A is an excellent standardized effect size for SCED research (Parker & Vannest, 2009). It is calculated by literally following its definition: by a brute force comparison of whether each data point in

condition is superior to each data point in condition A, and then calculating the percentage of all cases in which it is superior. Ruscio's A is therefore an effect size closely related to the Wilcoxon matched-pairs hypothesis test. BCA bootstrapping is then used to calculate 95% confidence intervals on A (Ruscio & Mullen, 2012).

It is useful to directly compare Ruscio's A to one particular SCED effect size metric: the Percent of Nonoverlapping Difference (PND; see REF for comparison of multiple different effect size metrics). PND is commonly used, computationally similar to Ruscio's A, and yet less robust that it. Specifically, PND counts the number of data points in condition B that are greater than the highest data point in condition A. This makes PND simple to calculate from a simple plot of the data. However, as such, it is highly sensitive to a single data point: the degree to which the highest data point in condition A is an outlier will influence the PND for that participant. Ruscio's A increases robustness by making comparisons between all data points in both conditions; effectively calculating a PND first for the highest data point in A, then for the second highest, and so on until a total probability of superiority value can be calculated.

Its one drawback of Rucsio's A that it suffers from a ceiling effect: if all data points in time point B are higher than time point A (i.e., Ruscio's A = 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. By using both the standardized and unstandardized effects sizes, the reader is given a rounded picture of the effect size. For example, an article might conclude that for a given participant "scores showed large increases after the intervention, P(A<B) = 1.00, 95% CI [0.92, 1.00], Mdn difference = 4.5."

Notionally, the confidence intervals on Ruscio's A could also be employed for decision making purposes rather than permuted *p* values, as they represent the confidence bounds of differences between the conditions. Of course, as mentioned above, this analytic choice should be made before data collection (e.g., in your study's preregistration) in order to limit researchers' degrees of freedom. Ruscio's A's CIs will not always agree with permuted *p*, particularly when the number of data points is very low in one or both conditions.

Meta-analysis of effect sizes. The above provide robust hypothesis test and effect size methods for individual participants in a SCED study. In order to pool results across participants, the SCED package also allows for the meta-analysis of Ruscio's A and Hedges' *g* effect sizes using the metafor R package (Viechtbauer, 2010), plus one unstandardized effect size. The meta-analysis of probability such values is still a matter of debate, and as such the SCED package opts to employ random effect model using the Maximum Likelihood estimator function over possible alternative methods (e.g., the conversion of probabilities to logits is problematic due to the plausible presence of values of 1.0). That is, although Ruscio's A for each participant is fully non-parametric, for ease of meta-analysis, the underlying effect is assumed to varies normally between participants. The results of the random effect meta-analysis provides both 95% confidence intervals on the meta effect size (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 on the basis of combining the CI with the observed heterogeneity in this effect size between participants).

Effect sizes for each participant, along with the meta-analyzed effect and both it confidence and credibility intervals are presented in a forest plot. Importantly, this forest plot employs asymmetric confidence intervals in order to correctly represent the confidence in estimated probabilities (i.e., Ruscio's A).

One unstandardized effect size is also produced by the package in order to provide an indication of the real world difference between the two conditions. This can be particularly useful when very large effect sizes

are observed, given Ruscio's A potential for ceiling effects noted above. In order to adhere to the SCED package's philosophy of employing highly robust and interpretable metrics, it employs the median median-difference between participants. That is, the median participant demonstrated this median difference between conditions.

Finally, this also provides information about the heterogeneity observed between participants (i.e., estimates of Q, I^2, and H^2). See the metafor package's documentation or other materials on meta-analysis for more details of these metrics. Briefly:

- Q and its p value: A measure of squared deviations. Depends on number of participants.
- I^2: Percentage of variation in the observed effects that is due to true heterogeneity as opposed to sampling variation. Metric is therefore a description of data in sample, not an underlying quality associated with the true effect. Can be thought of as analogous to the reliability of a scale, which also represents the percentage of true variance as portion of total variance. Range 0-100, lower values preferable. E.g., 0 = no variability in observed effects to be explained as a systematic influence (e.g., some due to some moderator) as its all just sampling variation. Does not depend on the effect size scale or the number of participants.
- H^2: [total variability / sampling variability], also expressed as [(true variability / sampling variability)/sampling variability]. Lower values preferable, i.e., refer to less true variation to be explained as systematic (e.g., due to some moderator).

Visual inspection and analysis

There is a long tradition of visual analysis of SCED data (see Lane & Gast, 2014 for an accessible primer). While this package produces several quantitative metrics, visual inspection remains to be important and therefore the package also contains methods to easily plot AB SCED data. The SCED package therefore produces a subplot for each participant to enable such analyses. 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 line fitted to each of the conditions.

In particular, these regression lines should be noted as important for diagnosing within-condition trends that may confound the interpretation of results. For example, if there is a trend towards improvement at baseline then differences between the conditions may not be due to the intervention. Improvement at baseline could also be due to method factors (e.g., repeated presentation of some measures could acts as a mini intervention itself) then this can be mitigated by a) staggering the intervention time between participants (e.g., a multiple baseline design, which is recommended either way) and b) by using a flexible number of time points at baseline and waiting for the last N data points to demonstrate a trend below a chosen value, and using only these data points in the analysis. Of course, these values should be chosen ahead of time and preregistered.

Of course, there are several other methods that could be used to diagnose trends within time points (e.g., significance of the linear regression slope, or significance of a non parametric linear regression slope such as the Theil-Sein slope). However, most suffer from issues such as violated parametric assumptions, arbitrary magnitude cutoffs, multiple testing corrections (if using flexible number of baseline data points), and/or paradoxical power implications (e.g., where more data increases power to detect problematic trends at baseline, punishing the researcher for collecting additional data collection to find a stable baseline). In light of this, the SCED package opts to employ simple Ordinary Least Squares regression slopes. The standardized beta coefficients of these regression lines are included in the tables produced by the quantitative analysis.

Example of how to conduct a SCED analysis in R

See the SCED.Rmd RMarkdown 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. The results table and SCED data plot should be included, along with the text output of the meta-analyzed effect size and heterogeneity metrics. The meta-analysis forest plot can optionally be included.

"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), also referred to as the Common Language Effect Size (McGraw & Wong, 1992), the Probability of Superiority (Ruscio & Mullen, 2012), Nonoverlap All Pairs (Parker & Vannest, 2009) and others, and 3) Hedges' *g*. The latter, a version of Cohen's *d* that is bias corrected small numbers 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 and computation: the probability that a randomly chosen data point in condition B is larger than a randomly chosen data point in condition A. 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 0.3]. [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 logits 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, $Q(\text{df} = 4) = 6.99$, $p = 0.14$, $I^2 = 47.36$, $H^2 = 1.90$. This suggests that participants responded to the intervention in a comparable manner and that results can be appropriately generalized across participants."

Author note

Institute of Psychology, University of Bern, Switzerland; ian.hussey@icloud.com

References

Acion, L., Peterson, J. J., Temple, S., & Arndt, S. (2006). Probabilistic index: an intuitive non-parametric approach to measuring the size of treatment effects. Statistics in Medicine, 25(4), 591–602. https://doi.org/10.1002/sim.2256

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects

Models Using lme4. Journal of Statistical Software, 67(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bulté, I., & Onghena, P. (2008). An R package for single-case randomization tests. Behavior Research Methods, 40(2), 467–478. https://doi.org/10.3758/BRM.40.2.467

Bulté, I., & Onghena, P. (2013). The Single-Case Data Analysis Package: Analysing Single-Case Experiments with R Software. Journal of Modern Applied Statistical Methods, 12(2), 450–478. https://doi.org/10.22237/jmasm/1383280020

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

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. Social Science & Medicine, 210, 2–21. https://doi.org/10.1016/j.socscimed.2017.12.005

Duan, N., Kravitz, R. L., & Schmid, C. H. (2013). Single-patient (n-of-1) trials: a pragmatic clinical decision methodology for patient-centered comparative effectiveness research. Journal of Clinical Epidemiology, 66(8 0), S21–S28. https://doi.org/10.1016/j.jclinepi.2013.04.006

Egan, J. P. (1975). Signal Detection Theory and ROC Analysis Academic Press Series in Cognition and Perception. London, UK: Academic Press.

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

Kirby, K. N., & Gerlanc, D. (2013). BootES: an R package for bootstrap confidence intervals on effect sizes. Behavior Research Methods, 45(4), 905–927. https://doi.org/10.3758/s13428-013-0330-5

Lane, J. D., & Gast, D. L. (2014). Visual analysis in single case experimental design studies: Brief review and guidelines. Neuropsychological Rehabilitation, 24(3–4), 445–463. https://doi.org/10.1080/09602011.2013.815636

Lawrence, M. A. (2016). ez: Easy Analysis and Visualization of Factorial Experiments. Retrieved from https://CRAN.R-project.org/package=ez

Manolov, R., & Moeyaert, M. (2017). How Can Single-Case Data Be Analyzed? Software Resources, Tutorial, and Reflections on Analysis , How Can Single-Case Data Be Analyzed? Software Resources, Tutorial, and Reflections on Analysis. Behavior Modification, 41(2), 179–228. https://doi.org/10.1177/0145445516664307

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

Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: A primer with examples. Human Brain Mapping, 15(1), 1–25. https://doi.org/10.1002/hbm.1058

Onghena, P., & Edgington, E. S. (1994). Randomization tests for restricted alternating treatments designs. Behaviour Research and Therapy, 32(7), 783–786. https://doi.org/10.1016/0005-7967(94)90036-1

Ottenbacher, K. J. (1990). Visual inspection of single-subject data: an empirical analysis. Mental Retardation, 28(5), 283–290.

Park, H.-S., Marascuilo, L., & Gaylord-Ross, R. (1990). Visual Inspection and Statistical Analysis in Single-Case Designs. The Journal of Experimental Education, 58(4), 311–320. https://doi.org/10.1080/00220973.1990.10806545

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

Parker, R. I., Vannest, K. J., & Davis, J. L. (2011). Effect Size in Single-Case Research: A Review of Nine Nonoverlap Techniques. Behavior Modification, 35(4), 303–322. https://doi.org/10.1177/0145445511399147

Ridgway, G. (2013, December 3). Illustrative effect sizes for sex differences. https://doi.org/10.6084/m9.figshare.866802.v1

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.

Sawilowsky, S. S. (2009). New Effect Size Rules of Thumb. Journal of Modern Applied Statistical Methods, 8(2), 597–599. https://doi.org/10.22237/jmasm/1257035100

Tate, R. L., Perdices, M., Rosenkoetter, U., McDonald, S., Togher, L., Shadish, W., ... Vohra, S. (2016). The Single-Case Reporting Guideline In BEhavioural Interventions (SCRIBE) 2016: Explanation and elaboration. Archives of Scientific Psychology, 4(1), 10–31. https://doi.org/10.1037/arc0000027

Tate, R. L., Perdices, M., Rosenkoetter, U., Shadish, W., Vohra, S., Barlow, D. H., ... Wilson, B. (2016). The Single-Case Reporting Guideline In BEhavioural Interventions (SCRIBE) 2016 Statement. Physical Therapy, 96(7), e1–e10. https://doi.org/10.2522/ptj.2016.96.7.e1

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

What Works Clearinghouse. (2010). Single-Case Design Technical Documentation. Retrieved from https://ies.ed.gov/ncee/wwc/Document/229

Wickham, H. (2010). A Layered Grammar of Graphics. Journal of Computational and Graphical Statistics, 19(1), 3–28. https://doi.org/10.1198/jcgs.2009.07098

Wickham, H. (2014). Tidy Data. Journal of Statistical Software, 59(10), 1–23.

Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Retrieved from http://ggplot2.org

Ximenes, V. M., Manolov, R., Solanas, A., & Quera, V. (2009). Factors Affecting Visual Inference in Single-Case Designs. The Spanish Journal of Psychology, 12(02), 823–832. https://doi.org/10.1017/S1138741600002195