ABkPowerCalculator: An App to Compute Power for Balanced $(AB)^k$ Single Case Experimental Designs

Prathiba Natesan Batley
ORCID: 0000-0002-5137-792X
Daiichi Sankyo

Madhav Thamaran
Texas A&M University

Larry Vernon Hedges
ORCID: 0000-0002-7531-0631
Northwestern University

Published in Multivariate Behavioral Research, 2023 59(2), 406–410.
doi:10.1080/00273171.2023.2261229..

The research reported here was supported by the Institute of Education Sciences, National Center for Education Research, through Grant R305D220052 to Northwestern University. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Abstract

Single case experimental designs are an important research design in behavioral and medical research. Although there are design standards prescribed by the What Works Clearinghouse for single case experimental designs, these standards do not include statistically derived power computations. Recently we derived the equations for computing power for $(AB)^k$ designs. However, these computations and the software code in R may not be accessible to applied researchers who are most likely to want to compute power for their studies. Therefore, we have developed an $(AB)^k$ power calculator Shiny App (https://abkpowercalculator.shinyapps.io/ABkpowercalculator/) that researchers can use with no software training. These power computations assume that the researcher would be interested in fitting multilevel models with autocorrelations or conduct similar analyses. The purpose of this software contribution is to briefly explain how power is derived for balanced $(AB)^k$ designs and to elaborate on how to use the Shiny App. The app works well on not just computers but mobile phones without installing the R program. We believe this can be a valuable tool for practitioners and applied researchers who want to plan their single case studies with sufficient power to detect appropriate effect sizes.

Keywords: Single case experimental designs; single case designs; power analysis; software; effect size

ABkPowerCalculator: An App to Compute Power for Balanced $(AB)^k$ Single Case Experimental Designs

Empirical studies require wise design choices that are guided by validity and statistical considerations. Single case experimental designs (SCEDs) are often useful in behavioral and medical research where randomization or collecting large samples might not be feasible or even appropriate. Some instances include psychiatric illnesses, autism spectrum disorders, behavioral disorders, comorbid conditions, and rare diseases (e.g., Au, Sauer-Zavala, King, Petrocchi, Barlow, & Litz, 2017; Hackett & Aafjes-van Doorn, 2019). Such small sample size scenarios often lead to possible underpowered studies. Therefore, an important consideration while planning SCEDs involves calculating power and sample size to procure sufficient data for conducting quantitative analyses to detect statistical significance when present.

Blackston Chapple, McGree, McDonald, and Nikles (2019) investigated power for aggregated N-of-1 trials (which are special cases of SCEDs) using simulations. However, this simulation study was restricted by the number of data conditions and the type of N-of-1 design was unspecified. Moreover, this does not provide a SCED researcher easily accessible information to compute the power for their own study. Bouwmeester and Jongerling (2020) designed a shiny app to compute power for ABAB, multiple baseline, and replicated ABAB and multiple baseline designs. However, this app does not consider a design-comparable effect size that account for small samples and autocorrelations and consider the within-case nature of SCEDs such as Hedges, Pustejovsky, and Shadish (2012, 2013). Several authors have estimated power for randomization tests for SCEDs using simulations which is yet another technique to estimate power (Levin et al., 2018, 2021; Michiels et al., 2020). Percha, Baskerville, Johnson, Dudley, and Zimmerman (2019) and Senn (2019) proposed power computations for N-of-1

designs. But, none of these have used design-comparable effect sizes, neither have they considered autocorrelations in their computations. Perhaps one of the most comprehensive studies in this area is the one by Wang and Schork (2019) which derived power based on distribution functions and considered autocorrelations. The only concern with this study is that they considered a within-subjects standardized mean difference effect size which is not appropriate for synthesizing SCED effects.

Addressing all these disadvantages, we derived power for balanced $(AB)^k$ designs with autocorrelations considering the design-comparable effect size (Hedges et al. 2012) in our recent paper (Hedges, Shadish, & Natesan Batley, 2022). The one drawback of this study is that it considers balanced designs with equal number of observations in the baseline and intervention phases. The k in an $(AB)^k$ design represents the number of repetitions of the AB phase. For instance, in an ABAB design k = 2. They showed that design comparable SCED effect size has the maximum impact on power followed by the number of subjects and then the number of phase reversals. That is, administering a treatment over 20 AB phases with 5 observations per phase yields more power than administering a treatment over 5 AB phases with 20 observations per phase although the total number of observations is the same for both cases. Previously mentioned power studies did not consider design-comparable effect size and in these studies autocorrelations had a large impact on power. However, in their computations, autocorrelations, the number of time-points per phase, and intraclass correlations had a smaller but non-negligible impact on power.

Although their study provides an essential framework for computing power for $(AB)^k$ designs, these computations require significant programming skills. In their previous paper they provided a link to an R code on their Github site. However, it is not common for SCED

practitioners who are subject experts in autism spectrum disorders, speech disorders, trauma research, medical research, or behavioral research to work through a moderately complex R program. Therefore, we created a Shiny App (https://abkpowercalculator.shinyapps.io/ABkpowercalculator/) that can be easily accessed and used by practitioners with little training. For more experienced programmers who want to work further on the code, the R code for this app is available on the Github site (https://github.com/prathiba-stat/ABk-power/blob/main/ABk-Shiny.R).

Model

Suppose that the Y_{ij} are normally distributed and that the data series for each individual i is weakly stationary within each phase with first order autocorrelation φ. Specifically, if there are n observations in each phase for each individual, the statistical model for the j^{th} observation which occurs in the p^{th} phase is

$$Y_{ij} = 0.5[1 + (-1)^{(p-1)}]\mu^C + 0.5\,[1 + (-1)^p]\,\mu^T + \eta_i + \varepsilon_{ij},\ \ i = 1, \ldots, m;$$

$$j = n(p-1) + 1, \ldots, pn;\ p = 1, \ldots, 2k.$$

In the expression ε_{ij} is normally distributed. The expressions in square brackets just assure that, in odd numbered phases (baseline phases), the coefficient of μ^C is one and the coefficient of μ^T is zero and that in even numbered phases (treatment phases), the coefficient of μ^T is one and the coefficient of μ^C is zero. The variable p is binary and takes the value of 1 for baseline and 2 for treatment phases. Thus, for example, the statistical model for the first (baseline) phase, where p = 1, is

$$Y_{ij} = \mu^C + \eta_i + \varepsilon_{ij},\ i = 1, \ldots, m;\ j = 1, \ldots, n,$$

and the statistical model for the second (treatment) phase, where p = 2, is

$$Y_{ij} = \mu^T + \eta_i + \varepsilon_{ij},\ i = 1, \ldots, m;\ j = n + 1, \ldots, 2n.$$

Here $\mu^T - \mu^C$ represents the shift between baseline and treatment periods. We assume that individuals are independent and that the individual effects η_i are independently normally distributed with variance τ^2. We assume that the ε_{ij} have variance σ^2 and first order autocorrelation φ within individuals. Therefore, the other assumption that the time series is weakly stationary implies that the covariance of Y_{ij} with $Y_{i(j+t)}$ depends only on t. The intraclass correlation is defined as the ratio of the variance between groups to the total variance. We refer the reader to the technical details of how this effect size is corrected for small samples and applied to ABAB type designs as described in Hedges et al. (2012). In essence, the design comparable effect size is a standardized mean difference effect size that corrects for small sample cases and considers autocorrelation and intraclass correlation in its computation. The interpretation of this effect size is like that of Cohen's d except that the effect sizes in SCEDs tend to be much larger than in other types of designs. Therefore, these effect sizes must be interpreted in accordance with what is considered a large effect for the substantive area in which the research is conducted.

The App

The parameters and symbols given in the Shiny App are described in Table 1. The app uses the packages shiny, psych, and shinyalert. Users may click on the link to the app which would open in their preferred internet browser. This app will require internet connection to function. Although the program runs in R, users do not have to install R on their systems to run this Shiny App. The program also functions on handheld devices such as smart mobile phones. The program requires the user to provide seven out of the eight parameters listed in Table 1 to compute the missing value. Usually, this would be power because researchers might be interested in calculating the power of their study. Alternately, while planning a study researchers

might want to input the power to estimate the number of subjects, number of phase repetitions, or the number of observations. The program computes the missing values within a few seconds. However, when users want to solve for k for large values of n and m (>15), the program might take up to 30 seconds, and for larger values of n and m (>25), the program might take up to 2 minutes to compute. It is of import to note that such extremely large values of phase repetitions are extremely uncommon in SCEDs. A recent study of 115 N-of-1/SCED studies found that the average number of subjects used in SCEDs since 2015 is 12.5 (Natesan Batley et al., Forthcoming). Consider a case where k = 2, autocorrelation = 0.5, and intraclass correlation = 0.5. To detect an effect of 0.75 with at least 80% power, it would require n = 35 observations per phase (a total of 140 observations per case over the 4 phases of the design) if m = 2 and n = 18 observations per phase (a total of 72 observations per case) if m = 3. Similarly, for a design with k=2, n=10, m=6, autocorrelation = 0.2, intraclass correlation = 0.4, an effect size of 0.45 yields a power of 0.8, whereas an effect size of 0.2 yields power of 0.24. The app has been rigorously tested by five independent researchers for its functionality and accuracy. Therefore, suffice it to say that this program functions well for any reasonable parameter value that would be used in SCEDs.

Other approaches

The technique proposed in this paper follows the trajectory of computing level and/or trend, intraclass correlation, autocorrelation, and design-comparable effect size for (AB)k SCEDs. However, what we have considered is only one approach that can be used in SCED data analysis. Visual analysis is still extremely popular, especially amongst applied researchers and considers more user-friendly computations and decision-making processes (Lane & Gast, 2014).

Manolov and Onghena (2018) proposed an approach that combines statistical approach with visual analysis. Much work has been done on individual aspects of SCEDs such as evaluating consistency within and across phases (Tanius, Manolov, & Onghena, 2021), Bayesian estimation of immediacy (Natesan & Hedges, 2017; Natesan Batley, Minka, and Hedges, 2020), etc.

It is necessary to remind the readers that statistical significance is sometimes not the aim of several popular SCED analyses such as visual analysis (Lane & Gast, 2014) and non-overlap indices (Maggin et al., 2019; Vannest et al., 2015). However, non-overlap indices that are commonly used in SCEDs suffer from drawbacks of being sensitive to outliers, being dependent on the scale of the outcome variable, and not having desirable distributional properties such as design comparable effect sizes. Sometimes finding statistical significance in SCEDs is purposely avoided as well (Branch, 2014, 2019; Perone, 1999).

Limitations

As discussed above, the app only computes power for $(AB)^k$ designs which are one type of SCEDs and multiple baseline designs that are widely used in SCEDs (Pustejovsky et al., 2019; Shadish & Sullivan, 2011; Smith, 2012) are not addressed here. The app can be used for only balanced designs, that is, equal time-points in each phase. Additionally, the derivations are based on asymptotic distributional assumptions. The power computations based on which the app is developed assume that the researcher would fit multilevel models with autocorrelations, compute design comparable effect sizes, or conduct similar analyses on the SCED data. We acknowledge that there are multiple methods of analyzing SCED data including visual analysis.

Although trend is evaluated by many researchers in SCEDs and is part of the WWC standards (2022), Natesan Batley and Hedges (2021) showed that estimating level, trend, and autocorrelations altogether in SCEDs leads to inaccurate estimates and increased Type-II errors.

They showed that researchers are better off estimating level and slope or a less accurate model of level and autocorrelation in their analyses. This study handles the latter case which is not the best performing model. Therefore, this is an area that requires further exploration. Autocorrelations in SCEDs have always remained contentious, especially in the presence of small sample data. Whether these need to be modeled and how they should best be modeled still needs further research in SCED data.

Discussion

The app presented in the present study is a useful tool to add to any SCED researcher's toolkit. SCEDs are becoming more important and necessary in an ever-increasing number of fields. For instance, one of the authors is currently leading a study on measuring clinically relevant, performance-based upper-limb prosthetic rehabilitation using extremely expensive prostheses. The expense of the prostheses limits the number of participants in the study. However, this prosthesis needs to be piloted before it can be manufactured on large scale for use with upper limb amputees. We computed the number of subjects and the other parameters of the study through these power calculations.

It might be difficult for researchers to choose values for intraclass correlation and autocorrelation for these computations. One suggestion is to examine the values found in literature and use them. In fact, to be safe, multiple such permutations of intraclass and autocorrelations could be used to derive a range of sample sizes and the final sample size decided based on the study parameters. It is also advisable to account for attrition of participants.

Researchers who are subject experts and not necessarily adept at programming in R or other statistical software can benefit from the use of this app. The app requires the input of 7 out of the 8 parameters. So, when a researcher wants to know the required sample size, number of

observations, and the number of phases for a given power this would require some back-and-forth computation. We are currently in the process of computing power for unbalanced and other types of single case experimental designs. We invite the research community to test and try the Shiny App and provide us feedback on its user-friendliness by emailing the first author.

References

Au, T. M., Sauer-Zavala, S., King, M. W., Petrocchi, N., Barlow, D. H., & Litz, B. T. (2017). Compassion-based therapy for trauma-related shame and posttraumatic stress: Initial evaluation using a multiple baseline design. *Behavior Therapy*, *48*(2), 207-221. https://doi.org/10.1016/j.beth.2016.11.012

Blackston, J. W., Chapple, A. G., McGree, J. M., McDonald, S., & Nikles, J. (2019). Comparison of aggregated n-of-1 trials with parallel and crossover randomized controlled trials using simulation studies. *Healthcare (Basel)*, *7*(4). https://doi.org/10.3390/healthcare7040137

Bouwmeester, S., & Jongerling, J. (2020). Power of a randomization test in a single case multiple baseline AB design. *PLoS One*, *15*(2), e0228355. https://doi.org/10.1371/journal.pone.0228355

Branch, M. (2014). Malignant side effects of null-hypothesis significance testing. *Theory & Psychology, 24*(2), 256−277. https://doi.org/10.1177/0959354314525282

Branch, M. N. (2019). The "reproducibility crisis:" Might the methods used frequently in behavior-analysis research help? *Perspectives on Behavior Science, 42*(1), 77–89. https://doi.org/10.1007/s40614-018-0158-5

Hackett, S. S., & Aafjes-van Doorn, K. (2019). Psychodynamic art psychotherapy for the treatment of aggression in an individual with antisocial personality disorder in a secure

forensic hospital: A single-case design study. *Psychotherapy*, *56*(2), 297-308. https://doi.org/10.1037/pst0000232

Hedges, L. V., Pustejovsky, J. E., & Shadish, W. R. (2012). A standardized mean difference effect size for single case designs. *Research Synthesis Methods*, *3*(3), 224-239. https://onlinelibrary.wiley.com/doi/10.1002/jrsm.1052

Hedges, L. V., Pustejovsky, J. E., & Shadish, W. R. (2013). A standardized mean difference effect size for multiple baseline designs across individuals. *Research Synthesis Methods*, *4*(4), 324-341. https://onlinelibrary.wiley.com/doi/10.1002/jrsm.1086

Hedges, L. V., Shadish, W. R., & Natesan Batley, P. (2022). Power analysis for single-case designs: Computations for (AB)(k) designs. *Behavioral Research Methods*. https://doi.org/10.3758/s13428-022-01971-9

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

Levin, J. R., Ferron, J. M., & Gafurov, B. S. (2018). Comparison of randomization-test procedures for single-case multiple-baseline designs. *Developmental Neurorehabilitation, 21*(5), 290-311. https://doi.org/10.1080/17518423.2016.1197708

Levin, J. R., Ferron, J. M., & Gafurov, B. S. (2021). Investigation of single-case multiple-baseline randomization tests of trend and variability. *Educational Psychology Review, 33*(2), 713–737. https://doi.org/10.1007/s10648-020-09549-7

Manolov, R., & Onghena, P. (2018). Analyzing data from single-case alternating treatments designs. *Psychological Methods, 23*(3), 480–504. https://doi.org/10.1037/met0000133

Michiels, B., Tanious, R., De, T. K., & Onghena, P. (2020). A randomization test wrapper for synthesizing single-case experiments using multilevel models: A Monte Carlo simulation study. *Behavior Research Methods, 52*(2), 654–666. https://doi.org/10.3758/s13428-019-01266-6

Maggin, D. M., Cook, B. G., & Cook, L. (2019). Making sense of single-case design effect sizes. *Learning Disabilities Research & Practice, 34*(3), 124–132. https://doi.org/10.1111/ldrp.12204

Natesan, P. & Hedges, L. V. (2017). Bayesian unknown change-point models to investigate immediacy in single case designs. *Psychological Methods, 22,* 743-759. DOI: 10.1037/met0000134.

Natesan Batley, P., Minka, T., & Hedges, L. V. (2020). Investigating immediacy in multiple phase-change single case experimental designs using a Bayesian unknown change-points model. *Behavior Research Methods, 52,* 1714-1728. DOI: https://doi.org/10.3758/s13428-020-01345-z.

Natesan Batley, P. & Hedges L.-V. (2021). Accurate Model vs. Accurate Estimates: A Study of Bayesian Single-Case Experimental Designs. *Behavior Research Methods, 53,* 1782-1798. https://doi.org/10.3758/s13428-020-01522-0.

Natesan Batley, P., McClure, E. B., Brewer, B., Contractor, A. A., Chin, S., Batley, N. J., & Hedges, L. V. (Forthcoming). Evidence and reporting standards in n-of-1 medical studies: A systematic review. *Translational Psychiatry*.

Percha, B., Baskerville, E. B., Johnson, M., Dudley, J. T., & Zimmerman, N. (2019). Designing robust n-of-1 studies for precision medicine: Simulation study and design

recommendations. *Journal of Medical Internet Research*, *21*(4), e12641. https://doi.org/10.2196/12641

Perone, M. (1999). Statistical inference in behavior analysis: Experimental control is better. *The Behavior Analyst, 22*(2), 109-116. https://doi.org/10.1007/BF03391988

Pustejovsky, J. E., Swan, D. M., & English, K. W. (2019). An examination of measurement procedures and characteristics of baseline outcome data in single-case research. *Behavior Modification*. Advance online publication. https://doi.org/10.1177/0145445519864264

Senn, S. (2019). Sample size considerations for n-of-1 trials. *Statistical Methods in Medical Research*, *28*(2), 372-383. https://doi.org/10.1177/0962280217726801

Shadish, W. R., & Sullivan, K. J. (2011). Characteristics of single-case designs used to assess intervention effects in 2008. *Behavior Research Methods, 43*(4), 971−980. https://doi.org/10.3758/s13428-011-0111-y

Smith, J. D. (2012). Single-case experimental designs: A systematic review of published research and current standards. *Psychological Methods, 17*(4), 510−550. https://doi.org/10.1037/a0029312

Tanious, R., Manolov, R., & Onghena, P. (2021). The assessment of consistency in single-case experiments: Beyond ABAB designs. *Behavior Modification*, *45*(4), 560-580.

Wang, Y., & Schork, N. J. (2019). Power and design issues in crossover-based n-of-1 clinical trials with fixed data collection periods. *Healthcare (Basel)*, *7*(3). https://doi.org/10.3390/healthcare7030084

What Works Clearinghouse. (2022). Procedures and Standards Handbook, Version 5.0. U.S. Department of Education, Institute of Education Sciences. Retrieved from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/Final_WWC-HandbookVer5.0-0-

508.pdf

Table 1: Symbols, parameters, and input values for the (AB)k Shiny App

Symbol	Parameter	Input
k	Number of phase repetitions	Positive integer>1
n	Number of observations per phase	Positive integer>1
m	Number of subjects	Positive integer>1
phi	Autocorrelation	$-0.99 \leq phi \leq 0.99$
rho	Intraclass correlation	$0 \leq rho \leq 0.99$
d	Design-comparable effect size	$0 \leq d$
alpha	Type-I error rate	$0 < alpha < 1$
Power	Power	$0 < power < 1$