Power analysis for single-case designs: Computations for (AB)k designs

Larry V. Hedges
Northwestern University

William R. Shadish
University of California, Merced

Prathiba Natesan Batley
University of Louisville

Accepted for Publication in Behavior Research Methods (2022)

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, 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

Currently the design standards for single case experimental designs (SCEDs) are based on validity considerations as prescribed by the What Works Clearinghouse. However, there is a need for design considerations such as power based on statistical analyses. We compute and derive power using computations for $(AB)^k$ designs with multiple cases which are common in SCEDs. Our computations show that effect size has the maximum impact on power followed by the number of subjects and then the number of phase reversals. An effect size of 0.75 or higher, at least one set of phase reversals (i.e., where k > 1), and at least 3 subjects showed high power. The latter two conditions agree with current standards about either having at least an ABAB design or a multiple baseline design with 3 subjects to meet design standards. An effect size of 0.75 or higher is not uncommon in SCEDs either. Autocorrelations, the number of time-points per phase, and intraclass correlations had a smaller but non-negligible impact on power. In sum, power analyses in the present study show that conditions to meet power requirements are not unreasonable in SCEDs. The software code to compute power is available on github for the use of the reader.

Power Analysis for Single Case Designs Based on Standardized Mean Difference Effect Sizes: Computations for $(AB)^k$ designs with Multiple Cases

In all empirical studies, wise design mandates that the data collection plan should provide the basis for inferences about the phenomenon under study that are as unambiguous as possible. When studies are conducted for the purpose of evaluating the efficacy of an intervention (in this paper we will use the word "treatment"), design focuses on organizing the data collection to ensure that unambiguous inferences about the treatment effect are possible. In experimental studies that use statistical hypothesis testing as a primary means of analysis, statistical power analysis plays an important role in design. Power analysis is often used to help the investigator determine whether the study, as planned, is sufficiently sensitive to detect effects that are expected (that is, whether the design has a sufficiently large probability of detecting a treatment effect of the size that is expected). Alternatively, sometimes power analysis is used to ensure that a study is likely to detect the smallest treatment effect deemed to be practically meaningful.

Power analysis plays a major role in designing experimental studies where the probability of detecting an effect (the statistical power) depends on several design parameters in complex ways. For example, in cluster randomized experiments (the most common design for randomized experiments used in education), statistical power depends on several factors that are under the control of the investigator: the number of clusters used, the sample size per cluster, and the significance level used. It also depends on several factors that are not under the control of the investigator: the treatment effect size, the ratio of between-cluster variation to total variation (also known as the intraclass

correlation), and the effectiveness of any covariates that are used to control variation at various levels of the design. Power analysis informs decisions about how to choose values of the design parameters that are under the control of the investigator (e.g., number of clusters and sample size per cluster) given the assumed values of the design parameters that are not under the control of the investigator (e.g., the intraclass correlation and the covariate-outcome correlations). For obvious reasons, agencies that fund experimental work involving statistics, such as the US Institute of Education Sciences (IES) or the National Institutes of Health (NIH), require power analyses to support claims about the sensitivity of the designs in studies proposed for funding.

Research using single case designs does not always use statistics as a primary mode of analysis. However, funding agencies typically expect that proposals for research should provide evidence that the designs chosen are sufficiently sensitive to detect the effects that treatments are expected to produce. As single case designs are increasingly used in research that will be evaluated by funding agencies like IES and NIH, and as statistical analyses for those designs become increasingly accepted, some principled means of addressing the issue of design sensitivity is needed.

One approach to the issue of design sensitivity builds on the work on statistical analysis of data from single case designs, namely statistical effect size measures (Hedges et al., 2012; Hedges et al., 2013). The focus of that work is not specifically the statistical analysis of single case designs, but the representation of effects obtained via measures of effect size that are in the same metric as those employed in between-subjects designs, so-called design-comparable effect sizes (Shadish et al., 2014). However, because the null hypothesis corresponds to an effect size of zero, the statistical properties of the effect size

estimate provide one method of statistical hypothesis testing and power analysis of the associated statistical test provides one means of assessing design sensitivity in single case designs. We would argue that this principled method of evaluating design sensitivity is useful even if the ultimate analysis does not use the associated hypothesis testing apparatus.

Analysis of the results of single subject designs has typically involved the visual search for functional relations between treatment assignment and outcome. That is, the study is designed so that each treatment (or baseline) phase is continued for enough measurements that the pattern of outcome values is clearly established. To establish functional relations, researchers often emphasize stability within treatment phases. Treatment effects are conceived as differences in these stable patterns between treatment and baseline phases.

Stability, however, can be conceptualized in several different ways. For example, the pattern could be one of fluctuation around a constant value with a common mean within a phase with a common residual variance within all phases. The pattern could also involve systematic increase or decrease across measurements in a phase, such as a linear or quadratic trend and a common residual variance within phases. Alternatively, the pattern could include a constant mean or a trend over measurements accompanied by systematic, increasing or decreasing residual variation around the trend. From this perspective, functional relations between treatment and outcome (what one would call treatment effects in between subject designs) are understood to be differences between the stable states established within treatment phases. The simplest pattern of stability, and the one that a given set of data has most information about, is one that involves

fluctuation around a constant value (a common mean with a common residual variance within phases). In this model of stability, treatment impacts correspond to shifts in the mean level of the outcome, although other models of stability are possible but not recommended for commonly seen SCED data conditions (see Natesan Batley & Hedges, 2021). We offered a statistical model in which the effect size parameter estimated corresponds to the standardized mean difference (Cohen's *d*), a well-known effect size parameter in between-subjects designs (Hedges, Pustejovsky, & Shadish, 2012).

In this article, we discuss power in the $(AB)^k$ design, the focus of Hedges, Pustejovsky, and Shadish (2012). In that design, A is typically a baseline phase, B is typically a treatment phase, and *k* indicates the number of times that the AB pair is repeated. For instance, $(AB)^2$ indicates an ABAB design in which the initial baseline phase (A) is followed by a treatment phase (B), then treatment is removed in a return to baseline (A), and the treatment is reintroduced (B). Shadish and Sullivan (2011) found that the $(AB)^k$ design was the second most frequently used design in their systematic sample of single-case designs in 2008.

Model for the $(AB)^k$ Design with *n* Observations Per Phase

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 be *n* observations in each phase for each individual, the statistical model for the j^{th} observation which occurs in the p^{th} phase is

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

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

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. 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. The assumption that the time series is weakly stationary implies that the covariance of Y_{ij} with $Y_{i(j+t)}$ depends only on t. We assume further that the ε_{ij} have variance σ^2 and first order autocorrelation φ within individuals. This autocorrelation model implies that the $2kn$ x $2kn$ covariance matrix of the errors within individuals for $2k$ phases is of the form given in notation N1

$$\frac{\sigma^2}{1-\phi^2}\begin{pmatrix} 1 & \phi & \ldots & \phi^{2kn-1} \\ \phi & 1 & \ldots & \phi^{2kn-2} \\ \vdots & \vdots & \vdots & \vdots \\ \phi^{2kn-1} & \phi^{2kn-2} & \ldots & 1 \end{pmatrix} \qquad \text{(N1)}$$

The Effect Size Parameter

It is conventional to assume that observations from different individuals are independent. Let us define the variance of observations within individuals within phases to be σ^2 and the variance of observations between individuals to be τ^2, so that the total variance of each observation is $\sigma^2 + \tau^2$. Define the mean of the observations in the

treatment phase by μ^T and the mean of the observations in the baseline (control) phase by μ^C. Under this model, define the effect size parameter as

$$\delta = \frac{\mu^T - \mu^C}{\sqrt{\sigma^2 + \tau^2}} . \quad (1)$$

This definition of the effect size is precisely the standardized mean difference (Cohen's *d*-index) that is widely used in between-subjects experiments. As we discuss below, this effect size parameter can be estimated from single case experiments as long as there are replications across individuals (that is $m > 1$). Note that the effect size parameter is the same in either the single subject design or a corresponding between-subjects design (see Hedges, Pustejovsky, and Shadish, 2012).

Estimation and Testing Hypotheses About δ

The numerator of the effect size is the unweighted difference between the means in the baseline and treatment phases, namely

$$\underline{D} = \frac{1}{mk}\sum_{i=1}^{m}\sum_{p=1}^{k}\left(\underline{Y}_{i\cdot}^{2p} - \underline{Y}_{i\cdot}^{2p-1}\right) = \frac{1}{mk}\sum_{i=1}^{m}\sum_{p=1}^{k}\left(\frac{1}{n}\sum_{j=n(2p-1)+1}^{2np} Y_{ij} - \frac{1}{n}\sum_{j=n(2p-2)+1}^{n(2p-1)} Y_{ij}\right) \quad (2)$$

where $\underline{Y}_{i\cdot}^{2p}$ is the mean of phase $2p$ and $\underline{Y}_{i\cdot}^{2p-1}$ is the mean for phase $2p - 1$ for individual *i*. Equation 2 assumes that there are equal number of observations in each phase.

The denominator of the effect size *S* is the square root of the variance across individuals at each timepoint but pooled across timepoints (and across phases). Thus, S^2 is defined as

$$S^2 = \frac{1}{2kn(m-1)} \sum_{i=1}^{m} \sum_{p=1}^{2k} \sum_{j=n(p-1)+1}^{np} (Y_{ij} - \underline{Y}_{.j})^2 \quad (3)$$

where $\underline{Y}_{.j}$ is the average across individuals of the j^{th} observations within individuals, given by

$$\underline{Y}_{.j} = \frac{1}{m} \sum_{i=1}^{m} Y_{ij}$$

The effect size estimate *ES* is therefore

$$ES = \frac{\bar{D}}{S}, \quad (4)$$

where $\bar{D}$ is given in (2) and S^2 is given in (3). The sampling distribution of this effect size is related to that of the noncentral t-distribution and was given in Hedges, Pustejovsky, and Shadish (2012). In equation 4, the expected value of $\bar{D}$ is the average of within-person contrasts. Let the constant *a* be the variance of $\bar{D}$, while *b* and *c* are the (standardized) expectation and variance of S^2. When $\sigma^2 \neq 0$ (so that $\rho \neq 1$), the statistic

$$t = \sqrt{\frac{b}{a}} \frac{\bar{D}}{S} \quad (5)$$

has the noncentral *t*-distribution with *h* degrees of freedom and noncentrality parameter λ that are given as[1]

$$h = 2b^2/c \quad (6)$$

$$\lambda = \sqrt{\frac{b}{a}} \delta . \quad (7)$$

[1] In the R program the non-central t-distribution is solved using the F distribution as a squared value of t

In equations 5, 6, and 7, the expressions for the constants *a, b*, and *c* depend on *k, n, m,* ϕ, σ, and τ and are given in the appendix to this paper. It turns out that these constants and the sampling distribution of the statistic *t* depend on σ and τ only through the ratio

$$\rho = \frac{\tau^2}{\tau^2 + \sigma^2}, \tag{8}$$

the proportion of the total variance that is between persons. Thus, ρ is a kind of intraclass correlation for single case designs.

Generally, *h* is a decreasing function of ρ, taking a maximum (which depends on φ, *k, n,* and *m*) when $\rho = 0$ and a minimum of $h = m - 1$ when $\rho = 1$, as expected. When $\rho = 1$, $\sigma = 0$, so the statistic *t* is just a one sample *t*-test on the baseline-treatment mean differences, which has $m - 1$ degrees of freedom. One interpretation of this behavior is that, when $\rho = 1$ (which implies $\sigma^2 = 0$), pooling the standard deviation across timepoints provides no more information about $\sigma^2 + \tau^2 = \tau^2$ than can be obtained from a single timepoint (because $\sigma^2 = 0$, observations do not vary across timepoints within individuals). However, when $\sigma^2 > 0$, pooling the standard deviation across timepoints does increase the information about $\sigma^2 + \tau^2$, so that the effective degrees of freedom are typically larger than $m - 1$. When both $\phi = 0$ and $\rho = 0$, $h = 2kn(m - 1)$ as expected since the *m* observations at any one of the $2kn$ timepoints (with $m - 1$ degrees of freedom at each timepoint) are independent of the observations at other timepoints.

When the null hypothesis is true $\lambda = 0$ so the statistic *t* has the central *t*-distribution with *v* degrees of freedom. The relation between *v* and the autocorrelation φ is more complex. Holding other parameters constant, *v* takes a maximum for a negative value of φ and decreases for larger or smaller values. In general, for moderate values of

φ (e.g., $-0.5 \leq \varphi \leq 0.5$), the dependence of v on φ is less pronounced than the dependence of v on ρ. Thus, a formal test of the hypothesis

$$H_0: \delta = 0$$

(conditional on ρ and φ) involves computing the statistic t given in (5) and rejecting H_0 if $|t| > c_{\alpha/2}$, where $c_{\alpha/2}$ is the two tailed level α critical value of the t-distribution with v degrees of freedom. Note that the degrees of freedom will typically be fractional, so that interpolation between tabled critical values (or computation of exact values based on non-integer degrees of freedom) will be necessary.

Power Analysis

The statistical power of the level α two-tailed test is

$$p = 1 - f(c_{\alpha/2} | \lambda, v) + f(-c_{\alpha/2} | \lambda, v) , \quad (9)$$

and the power of the level α one-tailed test is

$$p = 1 - f(c_{\alpha} | \lambda, v) \quad (10)$$

where $f(x | \lambda, v)$ is the cumulative distribution function of the noncentral t with noncentrality parameter λ and v degrees of freedom, and λ and v are given in (7) and (6). This distribution function is available in many statistical packages, including R, STATA, SAS, and SPSS.

Power values can also be computed using standard power tables (such as those in Cohen, 1977) for the one sample t-test. To use such tables, one typically enters the table on a row corresponding to the sample size and the column corresponding to the effect size and the value corresponding to that row and column is the power. Because those tables were designed for a slightly different purpose, it is necessary to enter the table with a synthetic sample size and a synthetic effect size to use such tables to compute the

statistical power of the test for treatment effects in single case designs. The synthetic total sample size is

$$N_{Synthetic} = h + 1 \tag{11}$$

and the synthetic effect size is

$$\delta_{Synthetic} = \sqrt{\frac{b}{a(h+1)}}\delta \quad , \tag{12}$$

where v is given by (6) above and a and b are given in the appendix. Note that interpolation between tabled values will usually be necessary because $N_{Synthetic}$ will usually not be an integer.

Generally, the parameters δ,φ, and ρ are not entirely under the control of the investigator. However, the number of subjects m and the number of observations per treatment period n are under the control of the investigator and they can be varied to ensure that the design has adequate sensitivity, given the values of δ, φ, and ρ. The situation is similar to that in the design of cluster randomized experiments, where power depends on the effect size, intraclass correlation, and covariate-outcome correlations, which are not under the control of the investigator, but are determined by the context of the experiment. The number of clusters randomized and the number of individuals per cluster *are* under the control of the investigator and can be varied to ensure that the design has adequate sensitivity to detect the effect size expected.

Expression (5), (6), (7), and (9) conceal the somewhat complicated relation between the design parameters m, n, δ, φ, and ρ and statistical power. The most obvious fact that follows from (7) is that power is an increasing function of the effect size δ and the number of cases m. Calculations using the results of this paper reveal other, less

obvious, relations. First power is a decreasing function of ρ, so that the larger the between-individual variance (as a fraction of the total variation), the lower the power.

Results

Table 1 gives the eta-squared values of an ANOVA where power is the dependent variable and the data conditions such as $k, m, n, \rho, \varphi, and\ d$. The italicized values show the total variance explained by planned contrasts. Effect size explains the most variation in power (36.69%) followed by the number of subjects (16.71%), and the number of phases (10.11%). Interestingly, the number of observations, the autocorrelations, and the intraclass correlation had a small effect on power with each having an effect size of 2.7% or lower. Having an effect size of 0.5 versus higher explained the most variation in power (27.09%) followed by 0.75 versus higher values (7.10%). However, effect sizes of 1 or larger had very little variability in power. Similarly, increasing the number of participants beyond 3 did not lead to much increase in power. Finally, the power for an $(AB)^k$ design where $k = 1$ was much smaller than the power for a design where $k \geq 2$ (8.82%). Beyond that, adding more phases did not lead to more power.

INSERT TABLE 1 ABOUT HERE

Figure 1 gives the statistical power as a function of ρ for φ = 0.1, 0.3, and 0.5 when m = 4, n = 4, k = 2, and δ = 0.75. These values for m and n are reasonably representative of the average single-case $(AB)^k$ design in the literature (Shadish & Sullivan, 2011); and $(AB)^k$ designs rarely have more pairs of phases than k = 2. Preliminary research suggests that effect sizes in single-case designs are typically used to investigate treatment that have relatively large effects (by the standards of between-subjects designs) and that effects are often larger than δ = 0.75 on average. Thus, the

calculations we report here may underestimate the typical power of single case designs that expect larger effects.

INSERT FIGURE 1 HERE

The relation between statistical power and autocorrelation is not monotonic across the entire range of possible values. Holding *k, n, m,* and ρ equal, power decreases towards a minimum as ϕ increases from 0, but at some point, begins to increase again as ϕ approaches 1. Although there is a decrease in power followed by increase in power, since the effect size of ϕ was only 2.1 we did not deem this to be practically significant enough to warrant an investigation. The location of the minimum depends primarily on ρ. Figure 2 gives the statistical power as a function of φ for $\rho = 0.2$, 0.5, and 0.8 when $m = 4$, $n = 4$, $k = 2$, and $\delta = 0.75$. While the shape of the functions is somewhat different for different values of ρ, they all seem to have a minimum in the vicinity of $\varphi = 0.6$ to 0.8.

INSERT FIGURE 2 HERE

Figures 1 and 2 provide some insight into the impact of design parameters that are not under the control of the investigator and will need to be imputed for the purposes of power analysis. Because the exact values of these parameters are unknown to the investigator planning a study, it is sensible (and likely considered essential by reviewers of research proposals) to estimate these parameters conservatively. That is, they should be imputed in a manner that is likely to err by underestimating, rather than overestimating, statistical power. The systematic behavior of power as a function of φ and empirical evidence from single case design studies (Shadish and Sullivan, 2011; Shadish, Rindskopf, Hedges, and Sullivan, 2013) suggests that using the value $\varphi = 0.5$ may be a sensible conservative default value for power analyses. Logical grounds dictate

that the between subject variation τ^2 is likely to be greater than within-subject-within-phase variation σ^2, an argument for which there is also some empirical support (Shadish & Sullivan, 2011; Shadish, Rindskopf, Hedges, & Sullivan, 2013). This would imply that $\rho = 0.5$ is a sensible conservative default value for power analyses. We urge the user to be skeptical of these default values and neither should be used if there is empirical or strong theoretical evidence about the values of these parameters when the study is being designed.

Turning to the variables under the control of the investigator, power is an increasing function of both m and n, but changes in m (the number of subjects) generally have a larger effect than corresponding changes in n (the number of observations per phase) when both are small. Figure 3 gives the statistical power as a function of m for $n = 3$, 6, and 9 where $k = 2$, $\rho = 0.5$, $\varphi = 0.5$, $\delta = 0.75$. In contrast Figure 4 gives the statistical power as a function of n for $m = 3$, 6, and 9 where $k = 2$, $\rho = 0.5$, $\varphi = 0.5$, $\delta = 0.75$. Comparing these two figures, the stronger dependence of power on m than on n is evident.

INSERT FIGURES 3 AND 4 HERE

Implications of Design Sensitivity for Standards for Single Subject Designs

Power calculations can shed some light on general recommendations about single case designs, like the standards that have been proposed by the US Institute of Education Sciences What Works Clearinghouse (WWC) (Kratochwill, et al., 2010). Those recommendations were made based on many factors, and design sensitivity was only one of them. However, it is interesting to evaluate the consequences of those recommendations for design sensitivity. In the context of $(AB)^k$ designs, their

recommendations imply at least $k = 2$ (to obtain 3 reversals) and $n = 3$ measurements per phase to meet standards with reservations or $n = 5$ measurements per phase to meet standards without reservations.

Consider first the implications for design sensitivity of the recommendation that $k = 2$ as opposed to $k = 1$ so that there are at least two reversals. Figure 5a gives the statistical power of the test for treatment effects as a function of δ for $k = 1$ and $k = 2$ where $n = 3$, $m = 3$, $\rho = 0.5$, and $\varphi = 0.5$. This figure shows that the statistical power is quite low even for effect sizes as large as $\delta = 1.5$, but that power is much higher for $k = 2$ than for $k = 1$. Figure 5b is analogous to Figure 5a, except that the number m of individuals is increased to $m = 5$. Comparing Figure 5a with Figure 5b when $k = 2$, we see that power is generally larger with $m = 5$ than with $m = 3$, and power to detect effect sizes greater than or equal to $\delta = .8$ is greater than 0.80 is achieved with when $m = 5$, but the power to detect effects of size $\delta = 0.75$ is still only 0.39. With $k = 2$, $n = 3$, $\rho = 0.5$, and $\varphi = 0.5$, a total of $m = 7$ cases is required to obtain a power of 0.80 to detect an effect of $\delta = 0.75$, but it would require $m = 15$ cases to do so when $k = 1$. Therefore, the requirement that studies have $k \geq 2$ is quite sensible from the perspective of design sensitivity and power.

INSERT FIGURES 5A AND 5B HERE

The WWC requires that there be at least three measurements per phase, that is $n \geq 3$, to meet standards with reservations and at least five measurements per phase ($n \geq 5$) to meet standards without reservations. Figure 6a illustrates the statistical power of the test for treatment effects as a function of δ for $n = 2$, 3, and 5 where $m = 2$, $\rho = 0.5$, and $\varphi = 0.5$. This figure shows that the statistical power is quite low even for effect sizes as large

as $\delta = 1$, but that power is higher for $n = 3$ than for $n = 2$. Figure 6b is analogous to Figure 6a, except that the number m of individuals is increased to $m = 5$. In Figure 6b, when $m = 5$, the difference between the power with $n = 2$ and $n = 3$ is smaller than in Figure 6a where $m = 3$. Comparing Figure 6a with Figure 6b when $n = 3$, we see that power is generally larger with $m = 5$ than with $m = 3$, and the difference between the power with $n = 2$ and $n = 3$ is smaller. Power to detect effect sizes of $\delta = 0.75$ is greater than 0.80 when $m = 5$ and $n = 5$, but the power to detect effects of size $\delta = 0.75$ is still only about 0.62 with $n = 2$ and 0.652 with $n = 3$. With $k = 2$, $n = 3$, $\rho = 0.5$, and $\varphi = 0.5$, a total of $m = 7$ cases is required to obtain a power of at least 0.80 to detect an effect of $\delta = 0.75$, but it would require 8 cases to do so when $n = 2$. Therefore, the requirement that studies have $n \geq 3$ is also sensible from the perspective of design sensitivity.

INSERT FIGURES 6A AND 6B HERE

The number of replications across cases (the value of m) has profound implications for design sensitivity. Figures 3 and 4 demonstrate that m has a greater effect on design sensitivity than does n. However, power is an increasing function of both m and n, and there are likely to be practical tradeoffs in the choice to increase one or the other. However, a case can be made, on design sensitivity grounds, that a sample size of $m = 2$ is too small to yield sensitive designs unless the effect size is exceptionally large. In the rest of this paragraph, consider a case where $k = 2$, $\rho = 0.5$, and $\varphi = 0.5$. To detect an effect of $\delta = 0.75$ 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 $m = 3$, but only

$n = 12$ observations per phase (a total of 48 observations per case) if $m = 4$, and $n = 8$ observations per phase (a total of 32 observations per case) if $m = 5$.

The findings mimic some of the earlier findings of simulation-based approaches in computing power for masked visual analysis (Ferron, Joo, & Levin, 2017), multilevel models (Shadish & Zuur, 2014), or other complex models (Heyvaert, Moeyaert, Verkempynck, et al., 2017; Natesan Batley & Hedges, 2021; Natesan Batley, Minka, & Hedges, 2020; Natesan & Hedges, 2017). All these studies, as expected show that more data means more information which means more power. However, not all types of data are the same. For instance, in this study we see that number of people can lead to more power than the number of observations.

The results given in this paper involve the assumption that the number of measurements in each phase for each subject is the same. This is analogous to the assumption of balance in experiments such as cluster randomized trials. This is typically a sensible assumption for assessing design sensitivity, but may be unrealistic in some studies, for example where the design is planned to give more observations during treatment phases than during baseline phases, or when $k > 2$ and its plan involves fewer observations in later phases of design. In such cases, the notation becomes considerably more complex as given in Appendix B.

Example

Suppose that we are contemplating an $(AB)^2$ design to investigate a treatment that is expected to have an effect size of $\delta = 0.75$ and wish to obtain a statistical power of at least 80% (0.80). We are unsure of the values of ρ and φ so we choose conservative values of $\rho = 0.5$ and $\varphi = 0.5$. We begin with the idea that we will observe $n = 3$ times in

each phase and consider a sample size of $m = 3$ cases. Substituting $k = 2$, $n = 3$, and $\varphi = 0.5$ into expressions (8), (9), and (10), we obtain $a = 0.1670$, $b = 1.6667$, and $c = 0.4571$. Then substituting the values of b and c, along with $\varphi = \rho = 0.5$ into (6), we obtain $h = 5.95$. Substituting the values of a and b, along with $m = 3$, and $\rho = 0.5$ into expression (7), we obtain $\lambda = 1.982$. Substituting the value $\lambda = 1.982$ and $h = 5.95$ into expression (11), we obtain a two-tailed statistical power of $p = 0.38$, which is less than the target value of $p = 0.80$. At this point, we could consider increasing n, the number of measurements per phase, or m, the number of cases. Computing the statistical power with $m = 3$ but $n = 7$, 8, and 9 yields power of $p = 0.47$, 0.51, and 0.55, respectively, still less than the target value. Computing the statistical power with $n = 3$ but $m = 5$ yields power of $p = 0.65$, and increasing n to 6, 7, and 8 with $m = 5$ yields power values of 0.78, 0.81, and 0.84. If we both increase m to $m = 6$ and increase n to $n = 5$, the statistical power becomes $p = 0.80$. Design choices that yield power at or above 0.80 might depend on costs and feasibility of a larger number of measurements per phase versus a larger number of cases, a decision that would be best made in the context of a particular investigation.

Conclusions

The planning of research designs involves many considerations. Design sensitivity (statistical power) is only one of them. We would never advocate that power or design sensitivity should be the *only* consideration in planning a research design. However sufficient design sensitivity is essential for statistical conclusion validity, and therefore should always be *one* consideration in planning research. We have provided one formal method of assessing design sensitivity of single case research. These methods

are consistent with recently developed methods for characterizing the effect size from single case designs. Thus, they provide a natural complement to statistical analysis procedures involving effect sizes.

We argue that these methods may also be useful in planning research even if researchers do not intend to use statistical methods to analyze their findings or numerical effect sizes to characterize their magnitude. One reason is that visual methods do not offer an analogue to numerical methods for assessing design sensitivity. While rigorous visual analyses have many advantages, it is difficult to believe that they would be substantially more sensitive than statistical methods. Therefore, in the absence of visual analogues to power analysis, these numerical methods may be useful substitutes as input for planning single case research studies.

One important caveat is that the effect sizes on which these methods are based address a specific kind of treatment effect: Shifts in the mean level of the outcome. The effect size measure, its associated significance test, and the power computations would not be relevant if a different kind of treatment effect were anticipated, such as a change in variation. However, a parallel method leading to numerical effect size measures, associated significance tests, and power analysis could be developed for treatment effects reflecting impacts on different stable patterns of outcome measurements. We are currently developing such methods. The power computations in the present study are only based on standardized mean difference effect sizes. However, there are several other effect sizes used in SCEDs for which power calculations would vary. This is an avenue for future research.

Power, as prescribed by the current study, can be computed only in designs with more than 1 subject. A typical ABAB type design uses only one participant, but investigations of treatments show that almost always ABAB type studies involve more than one subject. The power computations in the present study are only applicable to balanced designs, that is, with equal number of observations in each phase. This is a limitation considering that researchers might want to use longer phases for implementing their treatments and shorter baseline phases just to obtain consistency in data. We recognize that computing power using the codes given on github (https://github.com/prathiba-stat/ABk-power/blob/main/Power) might be challenging to applied researchers and having a graphical user interface (GUI) for this purpose would be helpful. We are also developing power computations for multiple baseline designs which are most used in SCEDs. These efforts are already underway. SCED data are often not intervally scaled and might be count or percentage data (Natesan Batley, Shukla Mehta, & Hitchcock, 2020), in which case the current power calculations might not hold to be very accurate. We are currently developing effect sizes for count data that are also computing power using Monte Carlo simulations (Natesan Batley & Hedges, under review). It might be interesting to explore the extent to which other violations of assumptions would affect power in $(AB)^k$ designs.

References

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences, 2nd Edition*. New York: Academic Press. https://doi.org/10.4324/9780203771587

Ferron, J. M., Joo, S.-H., & Levin, J. R. (2017). A Monte Carlo evaluation of masked visual analysis in response-guided versus fixed-criteria multiple-baseline designs. *Journal of Applied Behavior Analysis, 50*, 701-716. doi: 10.1002/jaba.410

Hedges, L. V. (2007). Effect sizes in cluster randomized designs. *Journal of Educational and Behavioral Statistics*, *32*, 341-370. https://doi.org/10.3102/1076998606298043

Hedges, L. V., Pustejovsky, J. E., & Shadish, W. R. (2012). A standardized mean difference effect size for single case designs. *Journal of Research Synthesis Methods, 3*, 224-239. 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. *Journal of Research Synthesis Methods, 4,* 324-341. DOI: 10.1002/jrsm.1086

Heyvaert, M., Moeyaert, M., Verkempynck, P., Van Den Noortgate, W., Vervloet, M., Ugille, M., & Onghena, P. (2017). Testing the inter- vention effect in single-case experiments: A Monte Carlo simulation study. *The Journal of Experimental Education, 85*(2), 175−196. https://doi.org/10.1080/00220973.2015.1123667

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M. & Shadish, W. R. (2010). Single-case designs technical documentation. Retrieved from What Works Clearinghouse website:

http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf DOI: https://doi.org/10.1177/0741932512452794

Natesan Batley, P. & Hedges, L. V. (2021). Accurate models vs accurate estimates: A simulation study of Bayesian single-case experimental designs. *Behavior Research Methods, 53,* 1782-1798. https://doi.org/10.3758/s13428-020-01522-0.

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. & Hedges, L. V. (under review). Design comparable Bayesian rate ratios for single case experimental designs with count data for unequal phase lengths.

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., Shukla Mehta, S. & Hitchcock, J. (2020). A Bayesian rate ratio effect size to quantify intervention effects for count data in single case experimental research. *Behavioral Disorders.* DOI: 10.1177/0198742920930704.

Pustejovsky, J. E., Hedges, L. V., & Shadish, W. L. (2014). Design-comparable effect sizes in multiple baseline designs: A general modeling framework. *Journal of Educational and Behavioral Statistics*, *39,* 368-393. https://doi.org/10.1177/0741932512452794

Shadish, W. R., Rindskopf, D. M., Hedges, L. V., & Sullivan, K. J. (2013). Bayesian estimates of autocorrelations in single-case designs, *Behavioral Research Methods*, *45*, 813-821.https://doi.org/10.1177/0741932512452794

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

Shadish, W.R., & Zuur, A.F. (2014). *Power Analysis for Negative Binomial Glmms for Single- Case Designs*. Society for Multivariate Experimental Psychology, Nashville, TN.

Technical Appendix

The theory leading to the distribution of *ES* is described in Hedges, Pustejovsky, and Shadish (2012) using a theorem from the appendix of Hedges (2007). It is somewhat simpler to explicate for balanced designs in matrix notation. Order the vector of $2kn$ observations for the i^{th} individual as

$$\mathbf{y}_i = (y_{i1}, \ldots, y_{i(2kn)}) \qquad \text{(A1)}$$

and define the $2kn$ x 1 contrast vector consisting of k repeats of the sequence $(\mathbf{1}_n', \text{-}\mathbf{1}_n')$ as

$$\mathbf{w} = (\mathbf{1}_n', \text{-}\mathbf{1}_n', \mathbf{1}_n', \ldots, \text{-}\mathbf{1}_n')', \qquad \text{(A2)}$$

where $\mathbf{1}_n$ is an n-dimensional column vector of 1's. Then the covariance matrix of $\mathbf{y}_i$ is

$$\mathbf{V}_i = \frac{\sigma^2}{1-\phi^2}\begin{pmatrix} 1 & \phi & \ldots & \phi^{2kn-1} \\ \phi & 1 & \ldots & \phi^{2kn-2} \\ \vdots & \vdots & \vdots & \vdots \\ \phi^{2kn-1} & \phi^{2kn-2} & \ldots & 1 \end{pmatrix} \qquad \text{(A3)}$$

Therefore, the within person contrast is $\mathbf{y}_i\mathbf{w}$, which has variance $\mathbf{w}'\mathbf{V}_i\mathbf{w}$ and the variance of $\bar{D}$ is

$$a = \mathbf{w}'\mathbf{V}_i\mathbf{w}/m\sigma^2 \qquad \text{(A4)}$$

The constants b and c are obtained from the expectation and variance of S^2. Let y_{ij} be the j^{th} measurement ($j = 1, \ldots, 2kn$) on the i^{th} person ($i = 1, \ldots, m$). Order the $2knm$ observations from all m individuals as

$$\mathbf{y} = (y_{11}, y_{21}, \ldots, y_{m1}, y_{12}, \ldots, y_{m2}, \ldots, y_{1(2knm)}, \ldots, y_{m(2kn)})', \qquad \text{(A5)}$$

then partition the $2knm$ x $2knm$ covariance matrix of $\mathbf{y}$ as

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \cdots & \boldsymbol{\Sigma}_{1(kn)} \\ \vdots & \ddots & \vdots \\ \boldsymbol{\Sigma}_{(kn)1} & \cdots & \boldsymbol{\Sigma}_{(kn)(kn)} \end{pmatrix}, \tag{A6}$$

where

$$\boldsymbol{\Sigma}_{ij} = \left(\tau^2 + \frac{\sigma^2 \phi^{|i-j|}}{1-\phi^2} \right) \mathbf{I}_m, \tag{A7}$$

τ^2 is the between subject variation, and $\mathbf{I}_m$ is an m x m identity matrix. We can write S^2 as a quadratic form in $\mathbf{y}$ as $\mathbf{y}'\mathbf{A}\mathbf{y}/2kn(m - 1)$, where $\mathbf{A}$ can be partitioned as

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \mathbf{A}_{(kn)(kn)} \end{pmatrix}, \tag{A8}$$

where $\mathbf{A}_{ii} = \mathbf{I}_m - \mathbf{1}_m\mathbf{1}_m/m$. then

$$b = \text{tr}(\mathbf{A}\boldsymbol{\Sigma})/\ 2kn(m - 1)(\sigma^2 + \tau^2) \tag{A9}$$

and

$$c = 2\text{tr}(\mathbf{A}\boldsymbol{\Sigma}\ \mathbf{A}\boldsymbol{\Sigma})/\ (2kn(m - 1)(\sigma^2 + \tau^2))^2, \tag{A10}$$

where tr($\mathbf{X}$) is the trace of a square matrix $\mathbf{X}$.

Appendix B

Power Computations in Unbalanced $(AB)^k$ Designs

For unbalanced designs power computations can be carried out by using the results from (Hedges et al., 2012; Hedges et al., 2013) Hedges, Pustejovsky, and Shadish (2012) in place of those given in this paper for balanced designs. Specifically, the noncentrality parameter λ given in expression (7) of this paper is replaced by

$$\lambda = \sqrt{\frac{E\{S^2\}}{V\{\bar{D}\}}}\delta,$$

where $V\{S^2\}$ is the variance of S^2 given in expression (26) and $V\{\bar{D}\}$ is given in expression (25) of Hedges, Pustejovsky, and Shadish (2012). Similarly, the degrees of freedom h in expression (6) of this paper is replaced by v in expression (28) of Hedges, Pustejovsky, and Shadish (2012). Once the substitutions are made, the power analysis in unbalanced designs proceeds in the same way as in the balanced case described in this paper.

Table 1: Eta-squared effect sizes in percentages with power as the dependent variable

Effect	Planned Contrast	Eta-sq
k (Number of phase repetitions)		10.11
	k = 1 vs 2, 3, 4	*8.82*
m (Number of subjects)		16.71
	m = 2 vs 3,...,10	*12.17*
n (Number of observations per phase)		0.59
phi (Autocorrelation)		2.1
rho (Intraclass correlation)		2.68
d (Effect size)		36.69
	d = 0.5 vs 0.75, 1, 1.5, 2, 2.5, 3	*27.09*
	d = 0.75 vs 1, 1.5, 2, 2.5, 4	*7.10*

Figure 1
Power of the $\alpha = 0.05$ two-tailed test as a function of ρ when $\varphi = 0.1$, 0.3, or 0.5 for effect size $\delta = 0.75$, when $k = 2$, $m = 4$, and $n = 4$

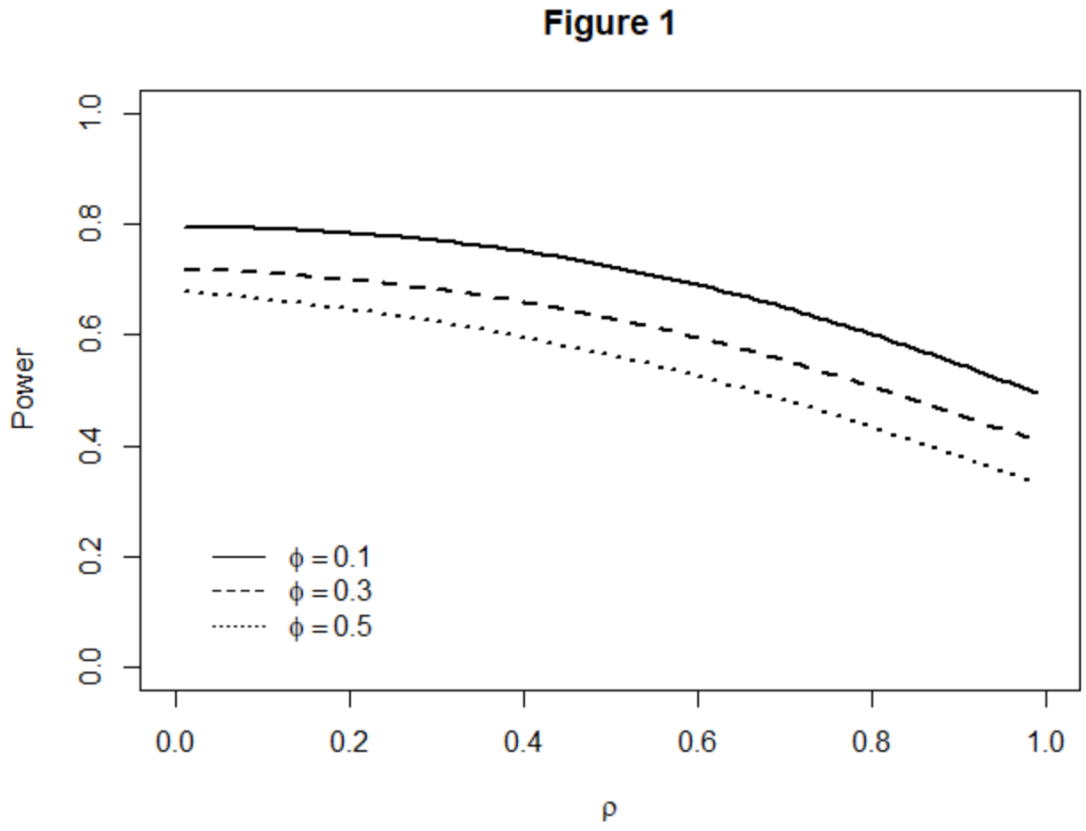

Figure 2
Power of the $\alpha = 0.05$ two-tailed test as a function of φ when $\rho = 0.2$, 0.5, or 0.8 for effect size $\delta = 0.75$, when $k = 2$, $m = 4$, and $n = 4$

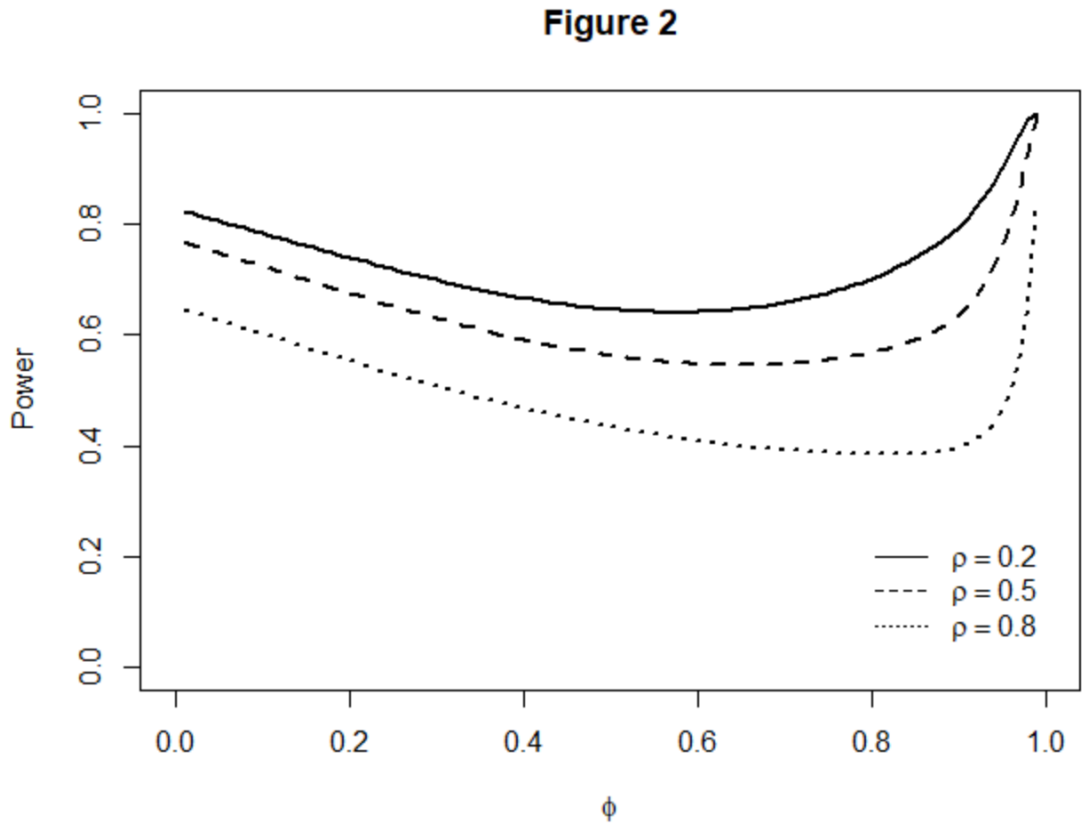

Figure 3
Power of the $\alpha = 0.05$ two-tailed test as a function of m when $n = 3$, 6, or 9 for effect size $\delta = 0.75$, when $k = 2$, $\varphi = 0.5$, and $\rho = 0.5$

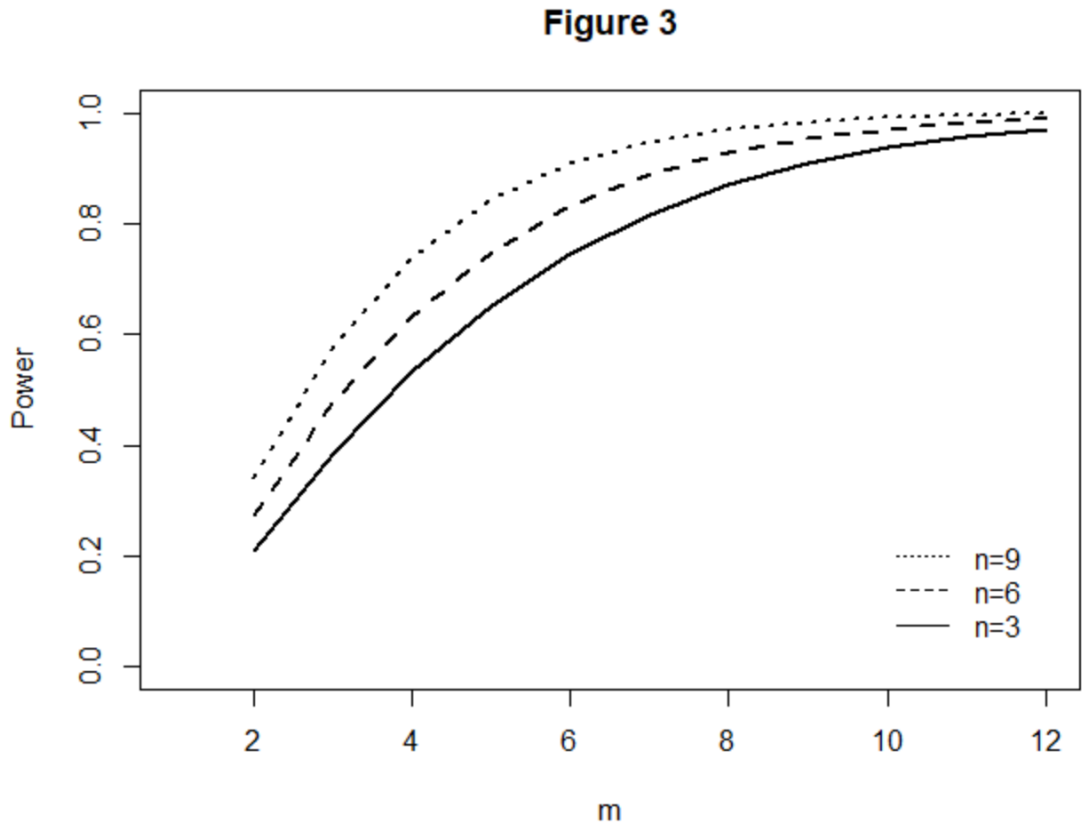

Figure 4
Power of the $\alpha = 0.05$ two-tailed test as a function of n when $m = 3$, 6, or 9 for effect size $\delta = 0.75$, when $k = 2$, $\varphi = 0.5$, and $\rho = 0.5$

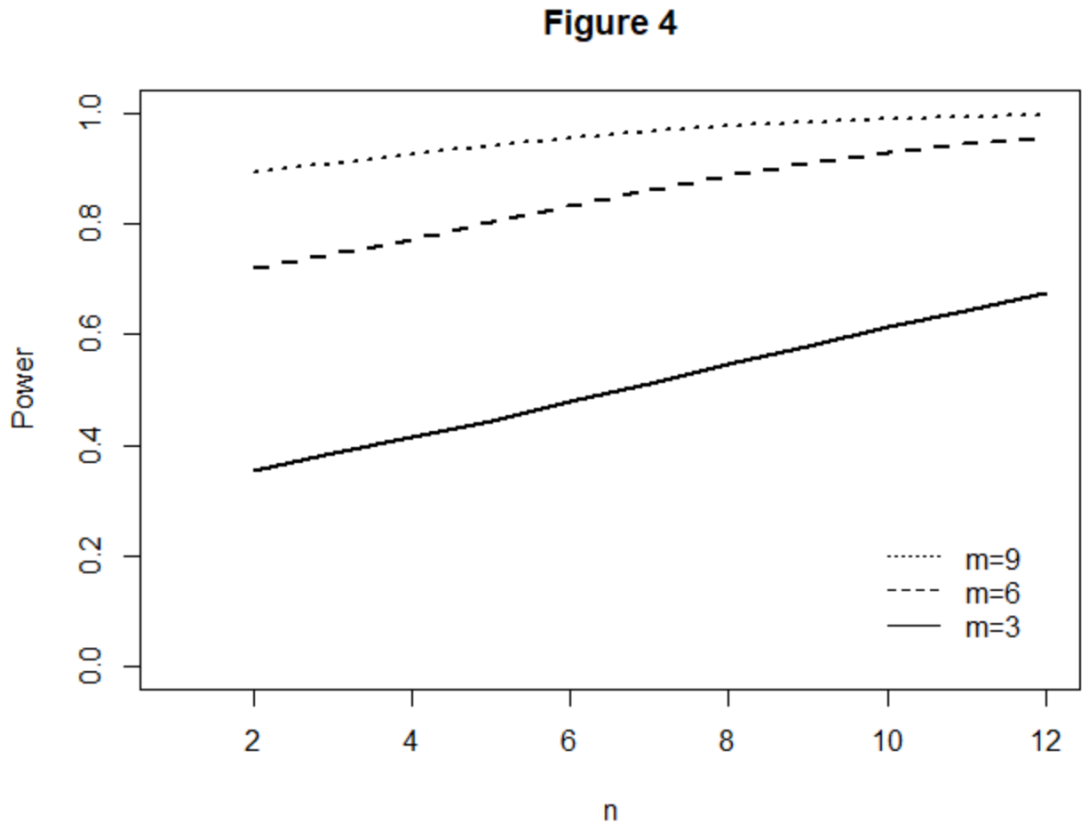

Figure 5a
Power of the $\alpha = 0.05$ two-tailed test as a function of effect size δ for $k = 1$, $k = 2$, and $k = 3$, when $n = 3$, $m = 3$, $\varphi = 0.5$, and $\rho = 0.5$

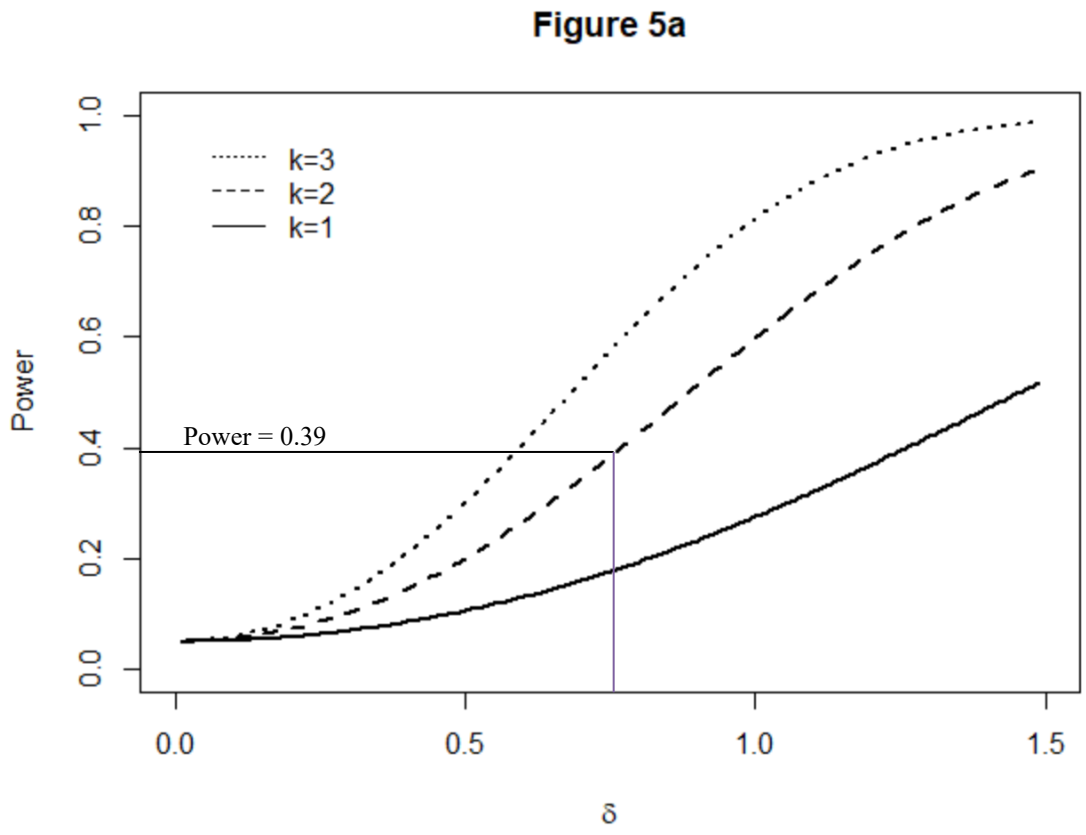

Figure 5b
Power of the $\alpha = 0.05$ two-tailed test as a function of effect size δ for $k = 1$, $k = 2$, and $k = 3$, when $n = 3$, $m = 5$, $\varphi = 0.75$, and $\rho = 0.5$

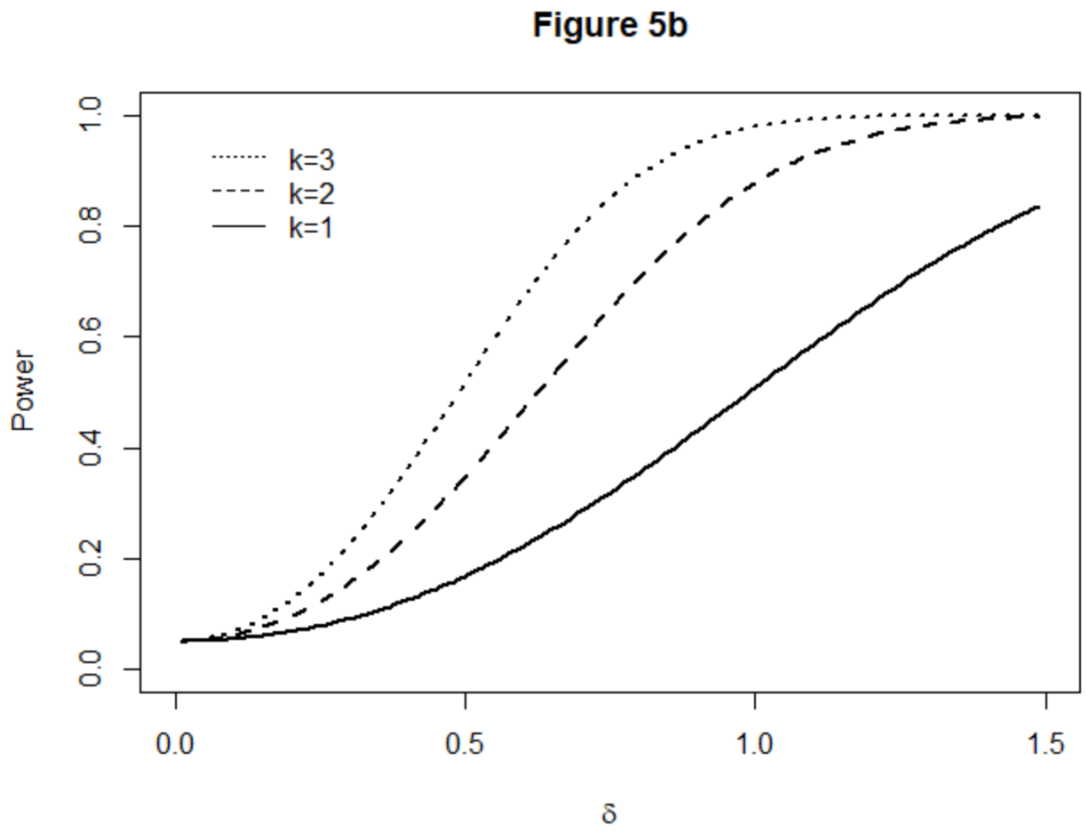

Figure 6a
Power of the $\alpha = 0.05$ two-tailed test as a function of effect size δ for $n = 2$, 3, and 5 when $m = 3$, $\varphi = 0.5$, and $\rho = 0.5$

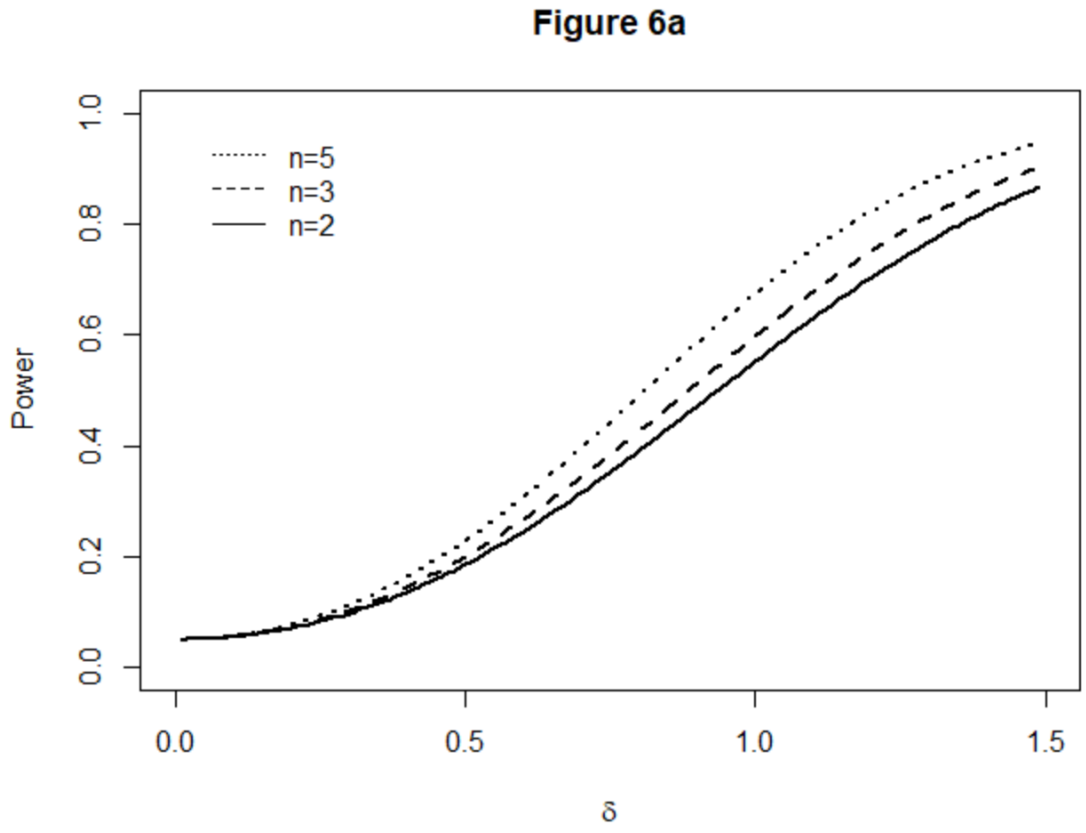

Figure 6b
Power of the $\alpha = 0.05$ two-tailed test as a function of effect size δ for $n = 2$, 3, and 5, when $m = 5$, $\varphi = 0.5$, and $\rho = 0.5$