Behav Res (2018) 50:557–575
DOI 10.3758/s13428-017-0885-7

The conditional power of randomization tests for single-case effect sizes in designs with randomized treatment order: A Monte Carlo simulation study

Bart Michiels[1] · **Mieke Heyvaert**[1] · **Patrick Onghena**[1]

Published online: 7 April 2017

Abstract The conditional power (CP) of the randomization test (RT) was investigated in a simulation study in which three different single-case effect size (ES) measures were used as the test statistics: the mean difference (MD), the percentage of nonoverlapping data (PND), and the nonoverlap of all pairs (NAP). Furthermore, we studied the effect of the experimental design on the RT's CP for three different single-case designs with rapid treatment alternation: the completely randomized design (CRD), the randomized block design (RBD), and the restricted randomized alternation design (RRAD). As a third goal, we evaluated the CP of the RT for three types of simulated data: data generated from a standard normal distribution, data generated from a uniform distribution, and data generated from a first-order autoregressive Gaussian process. The results showed that the MD and NAP perform very similarly in terms of CP, whereas the PND performs substantially worse. Furthermore, the RRAD yielded marginally higher power in the RT, followed by the CRD and then the RBD. Finally, the power of the RT was almost unaffected by the type of the simulated data. On the basis of the results of the simulation study, we recommend at least 20 measurement occasions for single-case designs with a randomized treatment order that are to be evaluated with an RT using a 5% significance level. Furthermore, we do not recommend use of the PND, because of its low power in the RT.

Keywords Single-case design · Randomization test · Statistical power · Nonoverlap effect sizes · Autocorrelation · Monte Carlo simulation study

Single-case experiments (SCEs) are designed experiments that include repeated measurements of a single entity (usually a person) for at least one dependent variable under different levels (i.e., treatments) of one or more independent variables (Barlow, Nock, & Hersen, 2009; Gast & Ledford, 2014; Kazdin, 2011; Kratochwill & Levin, 1992; Onghena, 2005).

Fields such as special education, school psychology, and clinical psychology are increasingly using SCEs to assess the efficacy of an intervention or treatment for a single subject (Alnahdi, 2015; Bowman-Perrott, Burke, de Marin, Zhang, & Davis, 2015; Hammond & Gast, 2010; Leong, Carter, & Stephenson, 2015; Moeller, Dattilo, & Rusch, 2015; Shadish & Sullivan, 2011; Smith, 2012; Swaminathan & Rogers, 2007). SCEs are also gaining in popularity in medical science (where they are often called "*N*-of-1 designs") to evaluate treatments for patients with, for instance, chronic pain or attention deficit hyperactivity disorder (Gabler, Duan, Vohra, & Kravitz, 2011). The recent development of guidelines for reporting the results of SCEs confirm the growing interest in these types of designs in the educational, behavioral, and health sciences (Shamseer et al., 2016; Tate, Togher, Perdices, McDonald, & Rosenkoetter, 2012).

Despite the growing popularity of SCEs, there is no broad consensus with respect to adequate data-analysis methods for these types of designs. As a result, a wide variety of methods is currently being used (often in combination with each other; Kratochwill et al., 2010; Maggin, O'Keeffe, & Johnson, 2011; Shadish, 2014). These methods can be broadly categorized in

Electronic supplementary material The online version of this article (doi:10.3758/s13428-017-0885-7) contains supplementary material, which is available to authorized users.

✉ Bart Michiels
bart.michiels@kuleuven.be

[1] Faculty of Psychology and Educational Sciences, KU Leuven–University of Leuven, Leuven, Belgium

two main approaches: visual analysis and statistical analysis (Heyvaert, Wendt, Van den Noortgate, & Onghena, 2015). Visual analysis consists of inspecting graphed SCE data for changes in level, overlap between phases, variability, trend, immediacy of the effect, and consistency of data patterns across similar phases (Horner, Swaminathan, Sugai, & Smolkowski, 2012). Statistical analysis methods for SCE data can be subdivided into three groups: effect size calculation, statistical modeling, and statistical inference. Effect size calculation refers to determining the size of the treatment effect by calculating formal effect size (ES) measures. Examples include mean difference measures (e.g., Busk & Serlin, 1992; Hedges, Pustejovsky, & Shadish, 2012), measures based on data nonoverlap between phases (e.g., Parker, Hagan-Burke, & Vannest, 2007; Parker & Vannest, 2009; Parker, Vannest, & Brown, 2009; Parker, Vannest, Davis, & Sauber, 2011), and regression-based measures (e.g., Allison & Gorman, 1993; Center, Skiba, & Casey, 1985–1986; Solanas, Manolov, & Onghena, 2010; Van den Noortgate & Onghena, 2003; White, Rusch, Kazdin, & Hartmann, 1989). In statistical modeling, the goal is to devise a statistical model that provides an adequate conceptualization of the data. Examples include multilevel modeling (Van den Noortgate & Onghena, 2003), structural equation modeling (Shadish, Rindskopf, & Hedges, 2008), and interrupted time series analysis (Borckardt & Nash, 2014; Gottman & Glass, 1978). Statistical inference refers to determining the statistical significance of ES measures through statistical hypothesis testing or to constructing confidence intervals for parameter estimates (Heyvaert, Wendt, Van den Noortgate, & Onghena, 2015; Michiels, Heyvaert, Meulders, & Onghena, 2017).

The present article deals with the inferential approach to evaluating treatment effects in single-case data. Inferential procedures can be parametric or nonparametric. However, parametric procedures such as statistical tests and confidence intervals based on t and F distributions are often not appropriate to analyze SCE data because the assumptions underlying these procedures (e.g., random sampling and more specific distributional assumptions) are often violated in many areas of behavioral research and particularly in single-case research (e.g., Adams & Anthony, 1996; Dugard, 2014; Edgington & Onghena, 2007; Ferron & Levin, 2014; Levin, Ferron, & Gafurov, 2014; Micceri, 1989). In contrast, nonparametric procedures do not make specific distributional assumptions about the data.

One of these nonparametric procedures, the randomization test (RT), has been proposed by some researchers as an appropriate statistical test to evaluate treatment effects in randomized SCEs (i.e., SCEs that include random assignment of measurement occasions to treatment conditions; e.g., Bulté & Onghena, 2008; Edgington, 1967; Heyvaert & Onghena, 2014; Levin, Ferron, & Kratochwill, 2012; Onghena, 1992; Onghena & Edgington, 1994, 2005). The RT is based on the random assignment model, which assumes that each experimental unit has been randomly assigned to one of the levels of the independent variable (similar to the way individual subjects are randomly assigned to treatment conditions in a between-subjects design; Kempthorne, 1955).[1] Furthermore, by randomly assigning measurement occasions to treatment conditions all known and unknown confounding variables can be controlled in a statistical way. Consequently, a potential statistically significant treatment effect can be attributed to the experimental manipulation. An alternative model, which is adopted by most parametric statistical tests, is the random sampling model. In this model, data are assumed to have been randomly sampled from a specific population of interest. Because the random assignment model does not make an assumption of random sampling, any statistical inference made under this model is conditional on the data that are analyzed (Keller, 2012).

A common practical problem in designing experiments is determining the number of observations that is required for the statistical tests to have sufficient power. The power of a statistical test is defined as the probability of rejecting a false null hypothesis. A power of 80% is generally accepted as the minimal requirement for a statistical test (Cohen, 1988, 1992). Power analysis can provide guidelines for the minimum number of observations that is required in order to detect an effect of a certain size with a certain probability. In an SCE the minimum number of observations refers to the minimum number of measurement occasions for the single case.

Apart from selecting the number of measurement occasions, the single-case researcher must also make other choices when designing a randomized SCE. More specifically, one must select a specific design, which determines the type of random assignment that is used in the SCE. In addition, the choice of an adequate ES measure is obviously important. All the aforementioned choices that are made when designing a randomized SCE have an effect on the power of the RT (Keller, 2012). It is thus extremely important for scientific practice to systematically investigate the effect of these factors on the power of the RT.

Several simulation studies concerning the power of the RT for different types of single-case designs and data patterns have already been performed (e.g., Ferron & Onghena, 1996; Ferron & Sentovich, 2002; Ferron & Ware, 1995; Heyvaert et al., 2017; Levin, Ferron, & Gafurov, 2014; Levin, Ferron, & Kratochwill, 2012; Manolov, Solanas, Bulté, & Onghena, 2010; Onghena, 1994). Although these simulation studies provide valuable information regarding the power of the RT in the context of analyzing SCEs, previous research has not yet systematically investigated one important determinant of the RT's power: the ES measure that is used as the test statistic. Furthermore, all previous simulation studies that examined the power of RTs for single-case designs have used a random sampling conceptualization of statistical

[1] We will use the term "assignment" to refer to a specific randomization of the condition labels in an SCE.

power, the so-called "unconditional power," although a random assignment conceptualization, the so-called "conditional power" is more consistent with the RT framework (Keller, 2012). With this article, we aim to fill both gaps.

With respect to the effect of the employed ES measure on the RT's power, we focused on nonoverlap effect size (NES) measures, which are currently receiving considerable attention from the single-case community as measures for quantifying treatment effects in SCEs (e.g., Heyvaert, Saenen, Campbell, Maes, & Onghena, 2014; Lenz, 2012; Wolery, Busick, Reichow, & Barton, 2010). NES measures are rooted either in the tradition of visually analyzing single-case data or in the tradition of nonparametric rank statistics, and assess the number of data points between conditions that do not overlap. Following the approach proposed by Heyvaert and Onghena (2014), we will use these NES measures as test statistics in an RT. More specifically, we included the percentage of nonoverlapping data (PND; Scruggs, Mastropieri, & Casto, 1987) and the nonoverlap of all pairs (NAP; Parker & Vannest, 2009) in our study.

The PND is the earliest published NES measure and the most widely used one (Maggin et al., 2011; Schlosser, Lee, & Wendt, 2008). The PND is calculated as the percentage of data points from the treatment condition that exceeds the single highest data point from the control condition (assuming that the treatment is intended to increase the dependent variable). To calculate the PND, one first identifies the highest data point from the control condition. Next, for each of the treatment condition data points, whether or not this data point exceeds the highest control-condition data point is recorded. The PND can take values from 0% to 100%, with 0% indicating complete data overlap, and 100% indicating complete data nonoverlap. Note that if the treatment is expected to decrease the scores on the dependent variable, the PND is calculated by comparing the treatment condition data points to the lowest control condition data point. Figure 1 illustrates the calculation of the PND for a hypothetical data set.

The NAP was introduced to meet the statistical limitations of the PND, and is calculated as the percentage of treatment data points exceeding each control data point by looking at all pairwise comparisons, with ties counting as a half point (Parker & Vannest, 2009). The NAP is equivalent to the Mann–Whitney U statistic and is defined from 0 to 1, with .50 indicating a null effect of the treatment (Mann & Whitney, 1947; Parker & Vannest, 2009). Figure 2 illustrates the calculation of the NAP for a hypothetical data set.

The second goal of this simulation study was to evaluate the power of the RT in a conditional power framework. In the previously cited simulation studies a random sampling model was used to generate the data for calculating the statistical power of the RT. Because the RT does not make an assumption of random sampling, evaluating the statistical power under a random sampling model does not do justice to the RT. As was demonstrated by Keller (2012), it is conceptually more appropriate to evaluate the statistical power of the RT by generating data using a compatible random assignment model. The resulting statistical power estimates are called "conditional power" estimates, because the estimates are conditional on a specific data set (see also Corcoran & Mehta, 2002; Gabriel & Hsu, 1983; Kempthorne, 1955; Kempthorne & Doerfler, 1969; Pesarin & De Martini, 2002; Pratt & Gibbons, 1981).[2] We elaborate this conditional power analysis approach in the Methods section and explain how this approach is combined with the three data generating processes that we used.

An additional goal of the present study was to investigate the effect of specific characteristics of the data on the power of the RT. Research has shown that data from single-case designs can contain autocorrelation (e.g., Shadish & Sullivan, 2011; Solomon, 2014). To account for this possibility we generated data that were not autocorrelated (independent standard normally distributed data) as well as data that contained strong positive autocorrelation (generated from a first-order autoregressive Gaussian process). In addition, we generated data from a uniform distribution (with a population standard deviation of 1) to evaluate the power of the RT in a situation in which classic distributional assumptions are severely violated.

The type of single-case design that is chosen to perform an SCE has important implications for the types of research questions that can be answered and the statistical power of the RT. For this reason we will now provide some information about the types of single-case designs we included in this simulation study and the types of research situations for which they are appropriate. The single-case designs that were used in this simulation study were all single-case alternation designs. Alternation designs are single-case designs in which rapid and frequent alternation of treatment conditions is possible. RTs for alternation designs are based on the random assignment of treatment conditions to measurement occasions (Onghena & Edgington, 2005). Although phase designs are used more often than alternation designs in practice (Shadish & Sullivan, 2011), we focused on alternation designs in this simulation study because they are more powerful than phase designs for SCEs (Onghena, 1994; Onghena & Edgington, 2005). The alternation designs we included were the completely randomized design (CRD), the randomized block design (RBD), and a restricted randomized alternation design (RRAD; Onghena, 1994, 2005). The CRD is the simplest alternation design (Edgington, 1967). In this design, treatment conditions are randomized solely with respect to the numbers of measurement occasions for each level of the independent variable. For example, the number of possible assignments for a hypothetical SCE with two conditions (A and B) with three

[2] This use of the term "conditional power" is standard in nonparametric statistics, but it should not be confused with the use of this term in the context of sequential clinical trials (see Lachin, 2005, for an overview of this alternative use of the term).

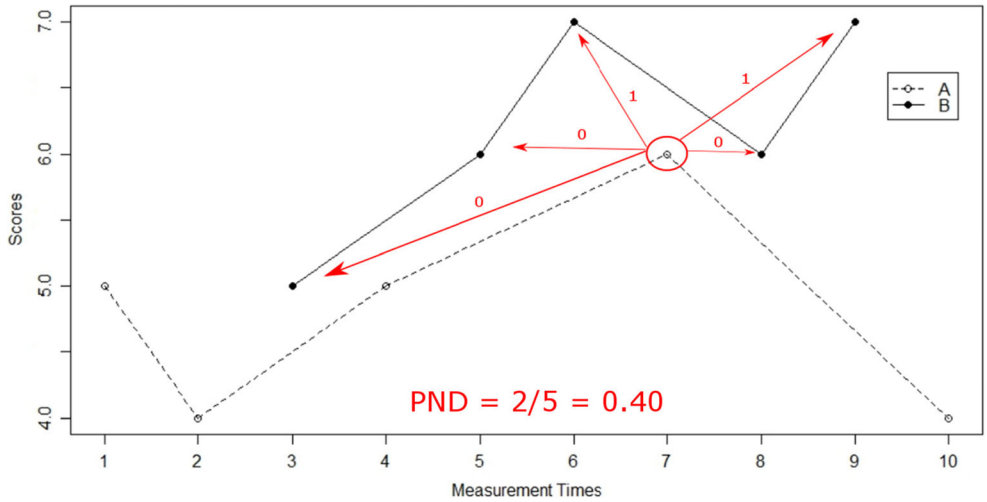

Fig. 1 Example of calculating the percentage of nonoverlapping data (PND) for a hypothetical data set. The dotted line represents the data from the control condition, and the full line represents the data from the treatment condition

measurement occasions per condition is given by $\binom{6}{3}$, which equals 20 possible assignments (Onghena, 2005):

AAABBB	BBBAAA
AABABB	BBABAA
AABBAB	BBAABA
AABBBA	BBAAAB
ABAABB	BABBAA
ABABAB	BABABA
ABABBA	BABAAB
ABBAAB	BAABBA
ABBABA	BAABAB
ABBBAA	BAAABB

This method of randomization is analogous to the random assignment of subjects to a balanced between-subjects design with two conditions. When certain assignments resulting from complete randomization are deemed undesirable for an SCE (e.g., AAAAABBBBB), other alternation designs can be derived from the CRD randomization scheme by imposing additional constraints on the method of randomization. For example, an RBD is obtained by grouping measurement occasions in pairs and randomizing the treatment order within each pair. An RBD for the same hypothetical SCE yields $2^3 = 8$ possible assignments (which are a subset of the set of CRD assignments):

ABABAB	BABABA
ABABBA	BABAAB
ABBAAB	BAABBA
ABBABA	BAABAB

RBDs can be used to counter the effect of time-related confounding variables on the dependent variable. For example, suppose a researcher wishes to conduct an SCE that evaluates the effect of a behavioral treatment on a depressed patient's feelings of negative affect. If the researcher knows that the level of negative affect of the patient can fluctuate from day to day, irrespective of the treatment, an RBD can be used to control for this confounding factor. Suppose the experiment consists of 10 days (i.e., ten blocks) where on each day the researcher administers both treatments (i.e., the control condition and the treatment condition) and records the patient's level of negative affect after each treatment (i.e., two negative affect scores per day). Because the sequence of conditions within a day (i.e., block) is determined randomly for every day, a potential significant treatment effect cannot be attributed to the day-to-day fluctuations in negative affect but only to the behavioral treatment.

When one wants to prevent the temporal clustering of treatments by ensuring that the randomization only allows a maximum number of successive measurement occasions to have the same treatment, one can use an RRAD (Onghena & Edgington, 1994). The RRAD yields a larger subset of the set of CRD assignments for a given SCE than the RBD. More specifically, an RRAD with a maximum number of two consecutive administrations of the same condition yields the following assignments for our hypothetical SCE:

AABABB	BBABAA
AABBAB	BBAABA
ABAABB	BABBAA
ABABAB	BABABA
ABABBA	BABAAB
ABBAAB	BAABBA
ABBABA	BAABAB

Note that the entire set of RBD assignments is present in the set of RRAD assignments.

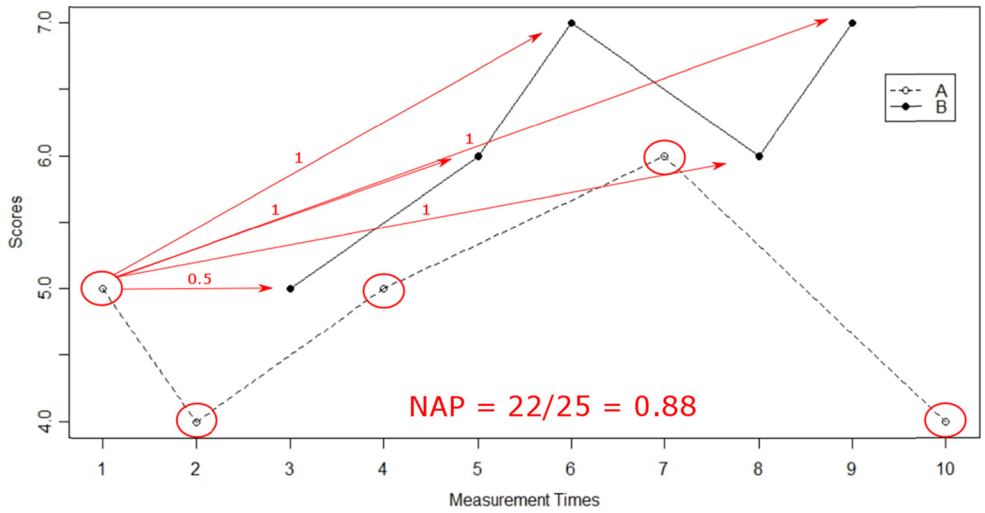

Fig. 2 Example of calculating the nonoverlap of all pairs (NAP) for a hypothetical data set. The dotted line represents the data from the control condition, and the full line represents the data from the treatment condition. The lines from every control condition data point to each of the treatment condition data points have only been drawn for the first control condition data point, so as not to clutter the graph

Method

The Method section contains three parts. The first part introduces the RT as a method of evaluating treatment effects in SCEs. The second part discusses how the conditional power of the RT is calculated. Finally, the third part details the design matrix of the simulation study.

Evaluating treatment effects in single-case experiments

Before explaining the way in which the conditional power of the RT is calculated, we will provide a worked example of the several steps that need to be taken to analyze a randomized SCE with an RT. Suppose we want to perform a randomized SCE consisting of four measurement occasions. Assume that the employed single-case design is balanced and completely randomized. The first step is made before executing the experiment and consists of listing all *permissible assignments* for the chosen experimental design. A permissible assignment is an assignment that adheres to the restrictions imposed by the chosen single-case design. In this example, the only restriction is that the design is balanced. When there are only two experimental conditions, this results in the following set of permissible assignments:

AABB
ABAB
ABBA
BAAB
BABA
BBAA

Second, one of the permissible assignments is randomly selected as the assignment to execute the actual experiment. Suppose the selected assignment is ABBA.

Third, one chooses an ES measure that is deemed adequate to answer the research question. This ES measure will be the test statistic of the RT. Suppose we choose the MD between the A and the B condition as the test statistic for this RT. Note that in order to test a two-sided alternative hypothesis, the test statistic must be sensitive to both directions of a possible effect. In this case we will use the absolute mean difference between the A and the B condition.

Suppose that the observed data are 2, 5, 4, and 3. For the selected assignment ABBA, this yields an observed test statistic of 2. As a fourth step, we calculate the test statistic for all permissible assignments:

$$\begin{aligned}
\text{AABB} &\Rightarrow |0| = 0\\
\text{ABAB} &\Rightarrow |-1| = 1\\
\text{ABBA} &\Rightarrow |2| = 2\\
\text{BAAB} &\Rightarrow |-2| = 2\\
\text{BABA} &\Rightarrow |1| = 1\\
\text{BBAA} &\Rightarrow |0| = 0
\end{aligned}$$

These values make up the randomization distribution. This collection of values is used as a reference distribution to calculate the statistical significance of the observed test statistic.

As a fifth step, the two-sided p value of the RT is calculated as the proportion of test statistics in the randomization distribution that are at least as extreme as the observed test statistic. When looking at the randomization distribution, we can see that two of the six permissible assignments lead to the same test statistic value as the observed test statistic, which results in a two-sided p value of 1/3. This p value provides a probabilistic statement of observing the data under the null hypothesis

that the conditions are unrelated to the data, and the validity of this statement is guaranteed by the randomization of the conditions. If that null hypothesis were true, then there is a probability of 1/3 to obtain a test statistic value as extreme as the one observed (Edgington & Onghena, 2007).

Note that this example was only chosen for didactical purposes as it is evident that an SCE with only four measurement occasions can never yield a p value that is smaller than any conventional significance level. Without performing any simulations, we can already infer that an SCE with only four measurement occasions has zero statistical power for all practical purposes.

The main advantages of the RT are that it makes no distributional assumptions and no assumption of random sampling. These advantages are important because it has been shown that the assumptions underlying parametric tests (e.g., random sampling or specific distributional assumptions) are doubtful in many domains of behavioral research and particularly for single-case research (e.g., Adams & Anthony, 1996; Dugard, 2014; Edgington & Onghena, 2007; Ferron & Levin, 2014; Levin et al., 2014; Micceri, 1989). Other advantages of the RT as compared to parametric tests are its flexibility with regard to the choice of test statistic and the choice of experimental design (Ferron & Sentovich, 2002; Onghena, 1992; Onghena & Edgington, 2005).

Power analysis in the random assignment model

The RT produces so-called "conditional inferences"—that is, inferences that are conditional on the observed data, just like Fisher's exact test is conditional on the marginal totals (Agresti, 1992; Krauth, 1988). Consequently, when investigating the power of the RT, it makes most sense to use this conditional framework too, and to calculate the so-called "conditional power" (i.e., the power of the RT for a specific data set). The advantage of this conceptualization is that the conditional power calculations are consistent with the random assignment model, which is also used for the validity of the RT, and that no assumption of random sampling is required. For the calculation of conditional power only an additional assumption about the treatment effect is necessary, just like one needs an assumption of the effect size parameter for the calculation of unconditional power.

If one would calculate the *unconditional* power of the RT, one would generate a large number of data sets (with fixed condition labels), sampled from a known distribution, and calculate the proportion of data sets that yield a p value smaller than or equal to a predefined significance level α. In contrast, to calculate the *conditional* power of the RT, one starts with a fixed set of scores that would be observed if the null hypothesis of no treatment effect is true (the "null scores"). Next one generates all possible randomizations of the condition labels, and constructs all possible data sets by pairing the null scores with the condition labels; null scores that are assigned to the treatment condition are transformed into observed scores containing the treatment effect. The conditional power is calculated as the proportion of those data sets that yield a p value smaller than or equal to α. Importantly, the unconditional power of the RT is defined in relation to the repeated random sampling of data sets from a known distribution whereas the conditional power of the RT is defined in relation to the repeated random assignment of condition labels to a specific set of null scores. Consequently for the calculation of conditional power, one does not need to make an assumption of random sampling. Note that this also implies that the resulting conditional power only pertains to that specific set of null scores.

To calculate either the conditional or the unconditional power of the RT one needs to make an assumption about the nature of the treatment effect. For the conditional power it means the specification of a specific functional relation between the null scores and the observed scores if a specific alternative hypothesis is true. The best known and most straightforward model in this respect is the unit-treatment additivity model (e.g., Cox & Reid, 2000; Hinkelmann & Kempthorne, 2008; Lehman, 1959; Welch & Gutierrez, 1988). This model describes the relation between the null scores and the observed scores as

$$X_i^B = X_i^A + \Delta$$

In this equation, X_i^B is the observed score of experimental unit i if i is assigned to the experimental condition B, X_i^A is the hypothetical score of i if i is assigned instead to the control condition A (i.e., the null score), and Δ is the constant additive effect of the treatment. If we assume that the null hypothesis is true, the equation above is reduced to

$$X_i^B = X_i^A$$

which implies that the observed score for experimental unit i is independent from the condition to which it is assigned. Note that the null scores are assumed to be known in order to calculate conditional power, just as the distributional form has to be known in order to calculate unconditional power.

The conditional power of the RT is determined by the significance level of the test, the number of observations, the size of the treatment effect, the employed test statistic, and the effect function (Keller, 2012). Whereas unconditional power is defined as the percentage of rejections of the null hypothesis across a given number of samples drawn from a certain population distribution, the conditional power is defined as the percentage of random assignments of treatment conditions to

experimental units that result in a rejection of the null hypothesis, given an assumed treatment effect (Kempthorne & Doerfler, 1969).

To calculate the exact conditional power of the RT for a specific data set, a few steps must be carried out. To begin, we must choose a single-case design, a number of observations, and a test statistic to use in the RT. Next, we generate one set of null scores for the chosen number of observations. We then obtain all permissible assignments of the employed randomization scheme for the chosen number of observations. If there are k permissible assignments, we then construct k different data sets from the null scores by adding the treatment effect to the null scores of the measurement occasions in the treatment condition. Next, we perform the RT for each of the k data sets from the previous step and record whether or not the null hypothesis is rejected at a pre-specified significance level. The exact conditional power is then defined as the overall proportion of rejected null hypotheses across the k RTs.

Notice that the RT is a computer-intensive method and that the calculation of the exact conditional power is "computer-intensive squared." If the number of observations rises, then k for each RT increases exponentially. For the exact conditional power, the RT is repeated k times, resulting in a total of k^2 calculations.

Table 1 illustrates the steps that are involved in calculating the exact conditional power of the RT.

In a random assignment framework with unit-treatment additivity, the conditional power of the RT is a function of the constant additive effect Δ. This implies that we can construct a conditional power curve for the null scores from

Table 1 Calculation of the RT's exact conditional power

Steps	Example				
1) Obtain a set of n (e.g., 4) null scores.	2, 3, 5, 3				
2) Generate all permissible assignments according to the employed single-case design (e.g., a balanced CRD).	A A B B				
	A B A B				
	A B B A				
	B A A B				
	B A B A				
	B B A A				
3) Apply the treatment effect (e.g., $\Delta = 1.5$) to the B data for each of the k permissible assignments resulting in k data sets.	k				
	1	A	A	B	B
		2	3	5 +**1.5**	3 +**1.5**
	2	A	B	A	B
		2	3 +**1.5**	5	3 +**1.5**
	3	A	B	B	A
		2	3 +**1.5**	5 +**1.5**	3
	4	B	A	A	B
		2 +**1.5**	3	5	3 +**1.5**
	5	B	A	B	A
		2 +**1.5**	3	5 +**1.5**	3
	6	B	B	A	A
		2 +**1.5**	3 +**1n5**	5	3
4) Execute the RT for each of the k data sets resulting in k p values.	k	p value			
	1	.33			
	2	1			
	3	.33			
	4	1			
	5	.33			
	6	1			
5) The exact conditional power of the RT is the proportion of the k p values that are smaller than or equal to, the chosen α level (e.g., $\alpha = 1/3$).	Half of the k p values are smaller than or equal to 1/3 so the exact conditional power is 50%.				

Table 4 by varying the value of Δ. Figure 3 displays the conditional power function for the set of null scores from Table 4 and for $\alpha = 1/3$.

For very small data sets such as in this example, it becomes apparent that the conditional power curve of the RT is actually a stepwise function. The function is stepwise because the conditional power is determined by the proportion of the k RTs that yield a p value smaller than or equal to the significance level α. If k is a small number, then only multiples of $1/k$ are possible conditional power values. For larger data sets, the number of possible steps is quite large so that the power curve becomes indistinguishable from a continuous function.

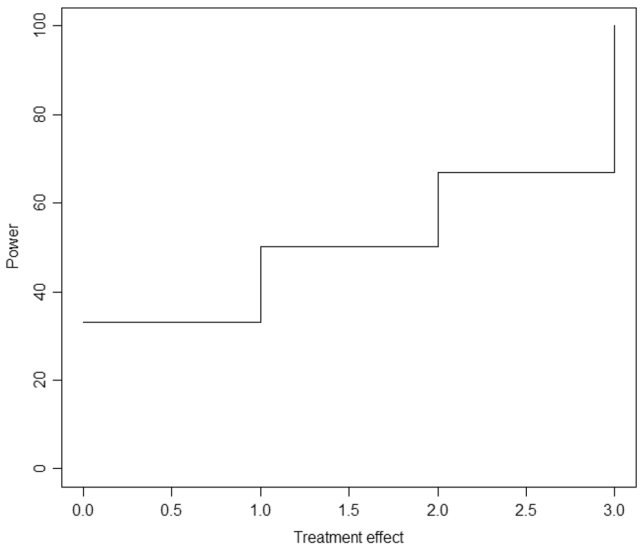

Fig. 3 Conditional power curve of the two-sided randomization test (RT) for $\alpha = 1/3$, in a completely randomized design using (2, 3, 5, 3) as the set of null scores

Design matrix of the simulation study

We manipulated five experimental factors in the present simulation study:

1. Characteristics of the data. To investigate the effect of different types of data on the conditional power of the RT, data were generated from a standard normal distribution and a uniform distribution (with a population standard deviation of 1). We selected the normal and uniform distributions because of their simplicity and ubiquity. Because both distributions were also used in the simulation study of Keller (2012), we could use his results as a benchmark. We added a first-order autoregressive model with Gaussian errors (AR1). The autoregressive parameter (AR) quantifies the autocorrelation in the data. We hypothesized that the presence of positive autocorrelation would have little influence in the selected single-case designs because of their fast alternation of the experimental conditions. Some pilot testing with small and medium levels of autocorrelation supported this hypothesis. For this reason (and in order to keep the number of experimental conditions manageable) we included only one, rather large, level of positive autocorrelation: AR = 0.6. Because the variance of an AR1 model is

$$\frac{\sigma_e^2}{1-AR^2}$$

and we sampled e from a standard normal distribution ($\sigma_e^2 = 1$) and AR = 0.6, the variance of the AR1 model is 1.5625. We only used three types of data (standard normal, uniform and AR1) to keep the simulation study feasible in terms of design and duration, and because we did not expect that the distributional shape would have a large impact on the power.

2. Test statistics used in the RT. Three different ES measures were used as the test statistic in the RT: the PND, the NAP, and the MD. The main reason for including the PND in this simulation study is that it is the most widely used NES (Maggin et al., 2011; Schlosser et al., 2008). As such, we believe it is of great importance to investigate PND's usability in statistical inferences. NAP was included because it was introduced to meet the statistical limitations of the PND (Parker & Vannest, 2009). To compare the performance of the selected NESs to a more generally accepted test statistic, we also included the mean difference (MD) in our simulation study. All tests statistics were formulated in a nondirectional way, so we only consider two-sided p values in this simulation study.
3. Designs. Three different single-case alternation designs were investigated: the CRD, the RBD, and the RRAD (see above for details). In this study, we limited our investigation to designs with two conditions (a control condition and a treatment condition). The CRD entails full random assignment of the condition labels with the only restriction that each condition must contain the same number of measurement occasions for each assignment. The RBD uses a form of randomization that groups measurement occasions into blocks of a certain size (we will use a block size of two observations) and then randomizes the measurement occasions within blocks. The RRAD uses full random assignment with the restriction that the maximal number of adjacent measurement occasions from the same condition is limited to a pre-specified value (Onghena & Edgington, 1994). In alignment with the What Works Clearinghouse (WWC) standards' recommendation to have at least three measurement occasions in a "phase" (Kratochwill et al., 2010), there could never be more than two adjacent measurement occasions from the same condition in the RRAD randomization scheme.

Note that all the designs were balanced designs (i.e., they contain the same number of measurement occasions in each condition).

4. Size of the treatment effect. Our choice of treatment ESs was guided by reported ESs in various domains of single-case research. ESs in single-case research are generally larger than in between-subjects research and are sometimes extremely high (Busk & Serlin, 1992). For example, Fabiano et al. (2009) performed meta-analyses of behavioral treatments for attention-deficit/hyperactivity disorder for various study designs and found average ESs of 0.83 and 3.78 for between-subjects studies and single-case studies, respectively. In a similar vein, two single-case meta-analyses concerning interventions for reducing challenging behavior in persons with intellectual disabilities resulted in average ESs of approximately 3 (Heyvaert, Maes, Van den Noortgate, Kuppens, & Onghena, 2012; Heyvaert, Saenen, Maes, & Onghena, 2014). With these results in mind, we included six levels of the treatment effect: 0, 0.5, 1, 1.5, 2, and 2.5. On the basis of pilot simulation testing, we set 2.5 as the maximum ES in our simulation study, because the conditional power for this ES was already 100% in almost all conditions. The size of the treatment effect in empirical single-case research can vary greatly depending on the specific domain and with the current selection of ESs we are able to cover the entire range of frequently found empirical ESs.
5. Number of measurement occasions. The selected numbers of measurement occasions to generate a complete data set were 12, 20, 30, and 40 measurement occasions and were chosen to cover the range of common series lengths in empirical research. For example, Ferron, Farmer, and Owens (2010) performed a survey that found average series lengths that ranged from 7 to 58 with a median of 24. A survey by Shadish and Sullivan (2011) found an average of 20 measurement occasions per individual time series. Note that the smallest amount of measurement occasions was selected to be 12 rather than 10 because of the fact that an RT using an RBD is unable to reach a 5% significance level for a data set with only 10 measurement occasions ($2^5 = 32$ possible assignments, and because we are considering two-sided tests, 1/16 is always larger than 1/20).

Crossing the levels of these five factors resulted in a total of $3 \times 3 \times 3 \times 4 \times 6 = 648$ conditions. For each condition, 100 null data sets were generated and the conditional power was averaged across these 100 replications. For each replication, the conditional power was calculated using Kempthorne and Doerfler's (1969) method. The significance level was set at 5%.

Despite rapid advancements in computing speed and exact algorithms, it is usually not feasible to calculate the exact conditional power of the RT due to the exponential increase of the computational demand when the number of observations grows larger (Keller, 2012). An alternative to exact computation is analytical approximation. For example, Gabriel and Hsu (1983) showed that their analytical approximation of the RT's exact conditional power only slightly overestimates the true power. However, their approximation shows larger biases for small numbers of observations or skewed treatment effects. In situations in which the number of observations is too large for exact computation but too small for precise analytic approximation, Monte Carlo approximation is a good alternative (Senchaudhuri, Mehta, & Patel, 1995). In this approach, only a random sample of all permissible assignments is used to calculate the conditional power of the RT. For a single RT it has been shown that the Monte Carlo RT produces valid p values (Edgington & Onghena, 2007). Furthermore, the accuracy of the random sampling can be increased to the desired level simply by increasing the number of random assignments that are drawn from the set of all permissible assignments (Senchaudhuri et al., 1995). Edgington (1969) pointed out that an efficient Monte Carlo RT can already be carried out with 1,000 random assignments. To keep the simulation study computationally manageable and without having to resort to analytical approximations, we used Monte Carlo sampling for the average conditional power calculation. More specifically, we selected 1,000 random assignments for each null sample, which resulted in 1,000 data sets for the conditional power calculation of each null sample. The conditional power for each null sample was calculated by performing the RT on each of these 1,000 data sets using 1,000 random assignments of the condition labels and by determining the proportion of p values that were equal to or smaller than .05. The average conditional power for each condition in the simulation study was obtained by averaging the conditional powers of the 100 null samples.

Results

To evaluate the effect of each experimental factor on the RTs conditional power, we plotted the main effect of each individual experimental factor while averaging the power across all other experimental factors. Figures 4 to 7 represent the main effects of ES measure, design, characteristics of the data, and number of measurement occasions on the average conditional power (averaged over 100 replications) with the size of the treatment effect plotted on the x-axis. Complete numerical results of the simulation study are displayed in Tables 2 to 4 in the Appendix.

Figure 4 shows that the MD and the NAP perform very similarly (an average difference of 1.74% in favor of the MD), whereas use of the PND as the test statistic in the RT yields substantially lower power. More specifically, the average

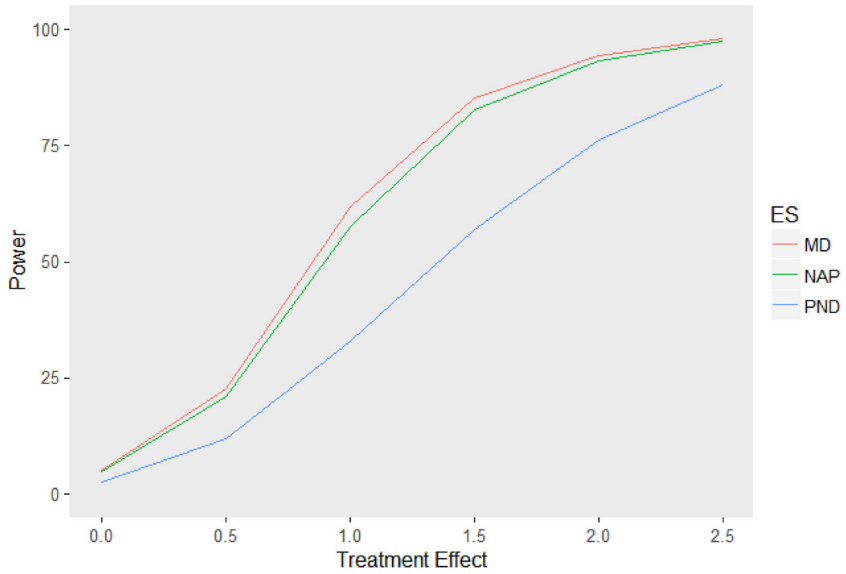

Fig. 4 Main effects of different effect size (ES) measures on the conditional power of the RT

power difference between the MD and the PND is 16.36% across the range of treatment ESs.

Figure 5 shows that the RRAD is the single-case design that on average yields the highest conditional power in the RT. More specifically, the average power advantage is 3.36% as compared to the CRD, and 6.39% as compared to the RBD.

Figure 6 reveals that there is only a very minimal, yet consistent, effect of the characteristics of the data on the conditional power of the RT. More specifically, the average difference between the conditions using the uniform distribution and the conditions using the standard normal distribution is 1.59%. Similarly, the average difference between the conditions using the standard normal distribution and the conditions using the AR1 model is 0.74%. These results show that the power of the RT is only slightly affected by extreme variations in the distributional characteristics of the data (cf. the data from a standard normal distribution vs. the data from a uniform distribution). Furthermore, the very minimal power difference between the conditions using the AR1 model and the conditions using the standard normal distribution indicates that the RT is almost not affected by strong positive autocorrelation in the data when single-case alternation designs are used.

Finally, Fig. 7 shows the effect of the number of observations on the conditional power of the RT. Note the sharply decelerating increases in conditional power when the number of observations is increased from 12 to 20 (an average increase of 15.65%), from 20 to 30 (an average increase of 8.69%), and from 30 to 40 (an average increase of 2.45%).

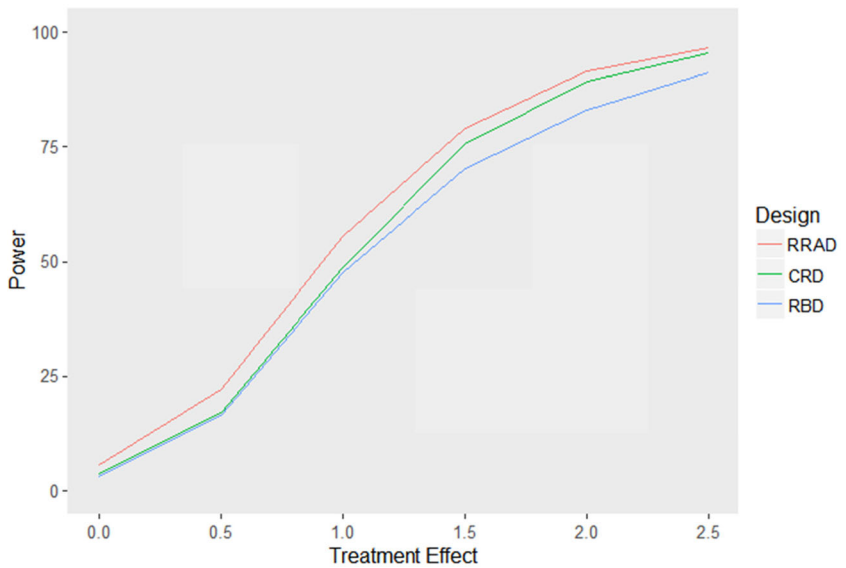

Fig. 5 Main effects of different designs on the conditional power of the RT

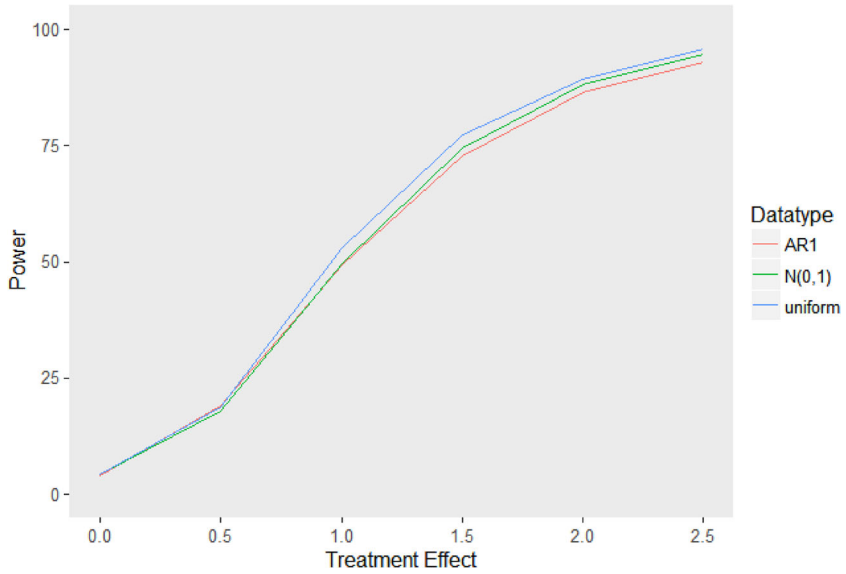

Fig. 6 Main effects of different characteristics of the data on the conditional power of the RT

Apart from visually analyzing the results of the simulation study, we also evaluated the results by looking at the variation between conditions using a multiway analysis of variance (ANOVA). More specifically we looked at the main effects of all experimental factors and two-way interactions effects that were deemed theoretically meaningful (an interaction effect between design and ES measure, between characteristics of the data and ES measure, and between characteristics of the data and design). We did not include higher-level interactions because they are difficult to interpret. For each evaluated effect, we calculated the proportion of explained variance in order to distinguish between the most important patterns in the results. All included effects of the ANOVA were significant at a significance level of .001. However, there are large discrepancies in the proportions of explained variance of the various effects. The size of the treatment effect by far has the largest proportion of explained variance (69.02%). This comes as no surprise because we used a wide range of values for this simulation factor (employing treatment ESs from 0 to 2.5). The number of measurement occasions has the second largest proportion of explained variance (6.9%). Furthermore, the choice of ES measure explains 3.34% of the variance, which is still considerable given the wide range of levels of the previously discussed factors. The other effects included in the ANOVA all explained less than 1% of the variance on their own. Nevertheless, they were all significant at the .001 level indicating that they have a significant effect on the conditional power of the RT. Furthermore, whereas the treatment effect is in practice not controlled by the researcher, and the number of measurement occasions is sometimes constrained by logistic

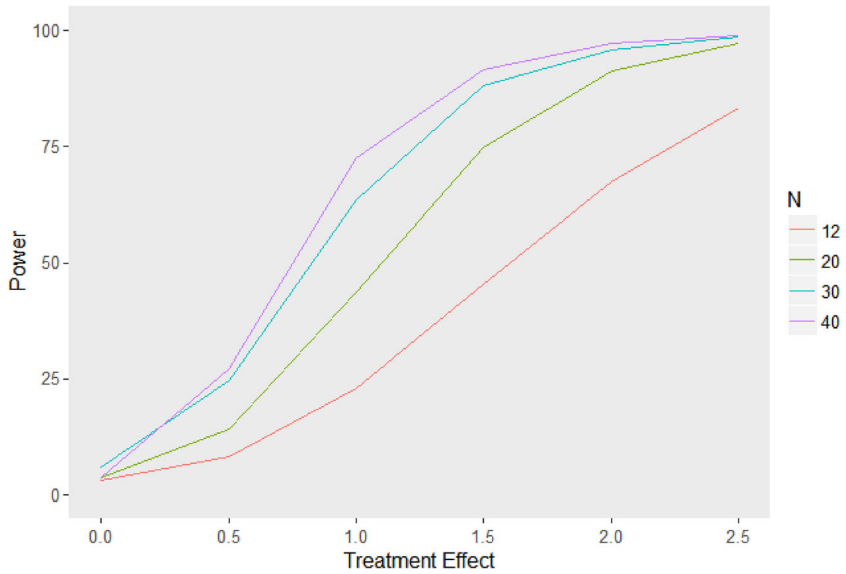

Fig. 7 Main effects of different numbers of observations on the conditional power of the RT

or financial considerations, the ES measure and the design can be chosen by the researcher. Hence it is reassuring to see that power can still be optimized by deliberate smart choices in the design phase of the study, given the constraints of the research context.

The main results from the simulation study can be summarized as follows:

- The MD and the NAP ES measures perform very similarly in terms of conditional power whereas the PND performs substantially worse.
- The RRAD is the single-case design that on average yields the highest conditional power.
- The conditional power of the RT is only minimally influenced by the characteristics of the null scores.

Discussion

In this article we investigated the conditional power of the RT using three different single-case ES measures (the MD, the NAP, and the PND) and three different randomized single-case designs with rapid treatment alternation (the CRD, the RRAD, and the RBD) for three types of simulated data (data from a standard normal distribution, data from a uniform distribution, and data from an autoregressive process with Gaussian errors) using a significance level of 5%.

The results were evaluated by visual analysis of power graphs and by decomposing the variance in the simulation results using ANOVA. The most important patterns in the results were identified by looking at the proportion of explained variance of each of the simulation factors.

With respect to the effect of the ES measure on the power of the RT, the results showed that the MD and the NAP perform very similarly whereas the PND performs substantially worse. This large discrepancy with the other two ES measures indicates that the PND has undesirable characteristics when it comes to evaluating intervention effects in SCEs. In fact, many authors have pointed out that the PND has two important limitations (e.g., Parker & Vannest, 2009; Shadish et al., 2008; Wolery et al., 2010). First, because only one control condition data point is used as a reference point to compare to the treatment data points, the PND is highly influenced by outliers. Second, the PND has a poor ability to discriminate between different magnitudes of a treatment effect. For example, when all treatment data points exceed the single highest control condition data point, it does not matter how large the nonoverlap is; the PND will always be 100%. Because of these limitations, authors have called for the abandonment of the PND in favor of other NES measures (e.g., Kratochwill et al., 2010; Parker & Vannest, 2009). In a similar vein, Campbell (2013, p. 24) stated, "I believe the PND methodology constitutes a first wave of SCD meta-analysis that is being followed by efforts to improve on inherent statistical limitations of PND." The results of our simulation study add to these criticisms, in the sense that the PND yields substantially less power in the RT than the NAP and the MD despite the absence of outliers in the simulated data and regardless of the size of the treatment effect. As such, we agree with the critics of PND that this measure should be abandoned in favor of more recent NES measures (such as the NAP).

With respect to the effect of the design on the conditional power of the RT, the results showed that the RRAD was the most powerful design, followed by the CRD and then the RBD. We argue that RRAD yields the highest power because it is a design that prevents the temporal clustering of measurement occasions but at the same time allows for a large number of assignments. In contrast, the CRD allows more temporal clustering than the RRAD whereas the RBD is more restrictive in terms of number of possible assignments than the RRAD. Although the choice of design also depends on feasibility and the concrete phenomenon that is studied, single-case researchers can take these findings with respect to the influence of the single-case experimental design on statistical power into consideration whenever they are designing an SCE that is to be evaluated using an RT.

The results of the simulation study also showed that the power of the RT is only minimally affected by the distributional characteristics of the null scores or the presence of strong positive autocorrelation. This is an important finding because it has been shown that single-case data often contain autocorrelation and violate classic distributional assumptions (e.g., Adams & Anthony, 1996; Dugard, 2014; Edgington & Onghena, 2007; Ferron & Levin, 2014; Levin et al., 2014; Micceri, 1989). However, we should note that the effect of autocorrelation on the power of the RT depends on the type of single-case design that is being used. For example, Ferron and Onghena (1996) evaluated the effect of autocorrelation on the power of the RT using a design that randomly assigns treatments to more extended phases. In this type of design, the researcher specifies a number of equally long phases and then randomly assigns a treatment condition to each of the phases. The results showed that positive autocorrelation increased the power of the RT for this type of single-case design. In contrast, Ferron and Ware (1995) showed that positive autocorrelation decreases the power of the RT for single-case phase designs that use random assignment of intervention points.

Apart from the size of the treatment effect, the number of measurement occasions is the second most important determinant of the conditional power of the RT. However, our results showed a sharply decelerating increase in conditional power when the number of measurement occasions was increased. In other words, the larger the number of measurement occasions becomes, the smaller the subsequent increase in power. This

information can be useful to researchers interested in SCEs for which the experiment is preferably as short as possible (e.g., if there are negative side effects of the experiment for the participant), in order to determine an optimal SCE length that balances between having sufficient statistical power and minimizing any discomforts for the participants.

We will now address some limitations to this simulation study. First of all, an obvious limitation is that the generalizability of our results is limited to the simulation conditions that were included. More in particular, we only considered null scores generated from continuous distributions. For some applications, other continuous or discrete distributions might be more relevant. We would expect that highly discrete distributions might compromise power because they give rise to many ties in both the data themselves and in the reference distribution. The power values in the present simulation study can therefore be considered as upper bounds. On the bright side, however, it should be noted that RTs always treat data in a discrete way and remain valid even if highly discrete (e.g., skewed dichotomous distributions) are used.

A second limitation is that the unit-treatment additivity model that we used to model the treatment effect in this simulation study conceptualizes the treatment effect as a difference in level and assumes no other effects. As a consequence, we only evaluated ES measures that are sensitive to differences in level. Nevertheless, we should mention that the unit-treatment additivity model is a generally accepted model in nonparametric statistics and a standard for classical evaluations of nonparametric methods (e.g., Cox & Reid, 2000; Hinkelmann & Kempthorne, 2008; Lehman, 1959; Welch & Gutierrez, 1988).

A third limitation of the present simulation study is that we used a Monte Carlo approach to approximate the exact conditional power. Although the Monte Carlo approach is computationally far more efficient, it introduces a random sampling (of assignments) error in the exact conditional power estimates. However, the magnitude of this error is a function of the number of random assignments that is used and can be determined analytically (see Edgington, 1969). More specifically, one can determine a confidence interval for the exact conditional power when it is approximated via Monte Carlo sampling. The bounds of a 99% confidence interval can be constructed using the following formula:

$$lower\ bound = \frac{(k-1)p - 2.58\sqrt{((k-1)pq)} + 1}{k}$$

$$upper\ bound = \frac{(k-1)p + 2.58\sqrt{((k-1)pq)} + 1}{k}$$

with k being the number of random assignments, p being the exact conditional power, q being $1–p$, and 2.58 being the 99.5th percentile of a standard normal distribution. For example, if the exact conditional power of a specific simulation condition is 80%, the 99% confidence interval of the Monte Carlo approximation with k = 1,000 is [77%; 83%]. Note that the Monte Carlo approach also causes the Type I error rates (i.e., the power when the treatment effect is zero) to slightly deviate from the specified significance level. When an exact RT is used (which uses all possible randomizations of the condition labels) the Type I error rate is always exactly equal to the specified significance level (Keller, 2012).

A generally accepted standard for sufficient statistical power is 80% (Cohen, 1988). When the power of a test is 80%, the probability of a Type II error (=1 – power) is 20%. If the conventional significance level of 5% is used, the ratio between the probability of a Type II error and a Type I error will then be four to one (20% to 5%). A power smaller than 80% would result in a too high a probability of a Type II error, whereas a power materially larger than 80% is likely to require data sets containing numbers of measurement occasions that could be unfeasible to collect (Cohen, 1992). To make a practical recommendation regarding the number of measurement occasions that yields a power of at least 80% in the RT using a significance level of 5% we must make an assumption about plausible ESs in single-case research. We previously mentioned that very high ESs of 3 or more are not uncommon in SCEs. In this case, the results of the simulation study with normally or uniformly distributed data show that an SCE with at least 20 measurement occasions is needed to obtain sufficient statistical power in the RT when used with randomized alternation designs and with a significance level of 5%. With other designs and highly discrete and skewed expected data sets even a larger number of measurements might be needed.

Finally, it is important to notice that the conditional power framework that we used in this simulation study contains an apparent paradox. On the one hand, the conditional power of a specific null data set only holds for that specific data set (i.e., it is conditional on the null scores). On the other hand, an a priori power analysis always requires knowledge, or an assumption, about the expected data in order to guide recommendations for the number of observations to be included in the experiment. Invoking distributional assumptions for the power analysis would go against the conceptual framework of conditional power as the latter makes no assumption of random sampling.

Similar to the approach taken by Keller (2012) we have tried to reconcile these two seemingly opposing requirements by calculating the conditional powers for a large number of null data sets (sampled from the probability distributions described in the methods section) and then by averaging these conditional powers. As such, the individual conditional powers are free of distributional assumptions, although the average conditional power for all the conditions reported in Tables 2, 3, and 4 of this article are dependent on the specific distributions from which the null data sets were generated.

In practice, a guesstimate of conditional power can be obtained by performing a small pilot study and using the pilot data to plan a subsequent larger data collection. The null scores in the unit-treatment additivity model can be reconstructed by calculating $\hat{\Delta}$ as the difference between the condition means and subtracting this difference from all the scores observed in the treatment condition:

$$X_i^A = X_i^B - \hat{\Delta}$$

Once we have null scores, we can calculate the exact conditional power for varying levels of Δ. Alternatively, different effect models can be explored, each leading to other null scores.

Collings and Hamilton (1988) and Hamilton and Collings (1991) also used this idea of a pilot study to determine the appropriate sample size on the basis of the distribution-free power of nonparametric tests for location shift. More specifically, the authors proposed to bootstrap the pilot data (i.e., draw random samples with replacement from the pilot data) to form a large number of bootstrap data sets. The proportion of data sets that yield a p value smaller than or equal to the significance level α is defined as the power estimate for the test. The authors showed empirically that their proposed method yields reliable results for estimating the power of the Wilcoxon two-sample test by means of a simulation study. Their method is appealing because the bootstrap technique allows for bootstrap data sets of any numbers of observations (smaller or larger than the number of observations in the pilot data set) and because no distributional assumptions are needed. In a similar vein we could use this bootstrap technique to generate a number of data sets from the pilot data. Next, we can calculate the conditional power for each of these data sets separately and obtain the average conditional power for all data sets together without additional distributional assumptions. The only additional assumption is that the distributional shape of the pilot data is indicative of the distributional shape of the finally collected data.

Suggestions for further research

To keep the computational burden of the simulation study manageable we had to limit the present study to the factors that were most relevant to answer our research questions. However, other interesting factors remain to be investigated in future research. These include: the use of various other ES measures as test statistics in the RT, the inclusion of other statistical distributions for generating data such as skewed, exponential or bimodal distributions, the inclusion of additional AR values for data generated with the AR1 model, and the inclusion of other time-series structures (e.g., moving average models). Because count data are used regularly as single-case outcome measures, future simulation studies could focus on the comparison between the power of the RT for data generated from discrete probability distributions and the power of the RT for data generated from continuous probability distributions. In addition future research could focus on investigating the power of the RT in unbalanced alternation designs as the single-case designs in this simulation study were all balanced designs.

In this study, we only modeled treatment effects that were defined as differences in level between experimental conditions. However, several single-case ES measures that look at trends exist (e.g., Tau-U, Parker et al., 2011; regression-based ES measures, Van den Noortgate & Onghena, 2003). Consequently, future research could focus on power analysis of the RT for ES measures that are sensitive to trend using simulated data that contain different types of trend effects. We previously mentioned that single-case phase designs are more frequently used in practice than single-case alternation designs. For this reason, future research could also investigate the conditional power of the RT for a variety of single-case phase designs.

Finally, another avenue for future research is to examine the theoretical relation between unconditional power and average conditional power as well as the possibility of obtaining distribution-free average conditional power using the bootstrap technique proposed by Collings and Hamilton (1988) and Hamilton and Collings (1991). For the latter it would be interesting to develop user-friendly statistical simulation software that assists in exploring the conditional power for a variety of distributional shapes and effect functions.

Conclusion

On the basis of the results of this simulation study, we would not recommend the use of the PND as an ES measure for the purpose of statistical inference using an RT in single-case alternation designs, because of its low statistical power. In contrast, the NAP yields power levels that are very similar to those of the MD, and as such provides a good alternative for researchers who want to use a nonoverlap measure to quantify the treatment effect in SCEs. With regard to the number of measurement occasions that are needed to ensure adequate statistical power in the RT, we recommend including at least 20 measurement occasions when working with alternation designs in the domain of single-case research and using a 5% significance level.

Author note We thank two anonymous reviewers for their valuable comments on a previous version of this article.

Compliance with ethical standards

Funding This research was funded by the Research Foundation–Flanders (FWO), Belgium (Grant ID G.0593.14 for B.M. and Grant ID 1242413N for M.H.).

Appendix: Results of the simulation study

Table 2 Conditional power, averaged over 100 replications for data sets generated from a standard normal distribution

Data: Standard normal distribution

Design		CRD			RBD			RRAD		
N	TE	Effect Size								
		MD	NAP	PND	MD	NAP	PND	MD	NAP	PND
12	0	4.82	3.97	1.08	3.26	3.22	0.18	5.65	4.86	2.04
	0.5	12.07	10.26	4.75	7.81	7.80	0.89	12.01	11.71	8.16
	1	32.74	31.45	15.63	21.57	23.39	4.75	33.96	31.90	20.47
	1.5	66.32	57.37	33.42	41.17	42.30	15.93	58.80	56.78	38.13
	2	89.11	84.60	60.09	64.81	59.70	29.88	85.62	80.41	64.00
	2.5	96.32	94.43	76.75	85.48	76.66	51.92	94.76	93.81	79.81
20	0	4.95	4.64	1.61	4.88	4.43	0.56	5.59	5.11	2.43
	0.5	17.66	17.28	8.93	16.75	13.70	4.45	18.69	16.54	11.61
	1	55.23	49.24	28.86	51.15	44.62	19.98	57.71	55.48	31.24
	1.5	90.54	86.64	55.71	83.77	81.41	42.15	88.64	85.61	58.29
	2	98.86	97.94	81.21	96.14	96.40	73.25	98.70	97.56	82.58
	2.5	99.98	99.98	94.21	99.91	99.62	89.43	99.94	99.76	92.65
30	0	4.94	4.90	1.80	4.97	4.78	0.96	7.01	7.77	17.55
	0.5	26.28	24.70	11.88	24.62	24.54	8.18	31.60	28.78	35.19
	1	74.67	71.97	34.06	69.40	67.12	31.44	72.73	74.60	56.14
	1.5	97.72	97.89	66.51	97.60	96.27	60.50	96.29	97.26	78.37
	2	99.97	99.93	88.20	99.82	99.79	82.21	99.89	99.81	91.79
	2.5	100	100	96.11	100	100	96.31	100	100	98.00
40	0	4.98	4.79	1.78	5.02	4.80	1.03	5.10	5.10	2.54
	0.5	33.74	31.63	11.93	31.51	29.36	9.66	31.99	31.99	15.09
	1	86.85	84.76	38.36	84.36	86.16	31.64	85.95	84.77	42.19
	1.5	99.60	99.32	75.35	99.51	99.32	65.10	99.36	99.44	76.13
	2	100	99.99	91.49	100	99.98	89.09	100	100	91.83
	2.5	100	100	98.40	100	100	96.10	100	100	97.77

N: number of measurement occasions; TE: Treatment Effect; CRD: Completely Randomized Design; RBD: Randomized Block Design; RRAD: Restricted Randomized Alternation Design; MD: Mean Difference; NAP: Nonoverlap of All Pairs; PND: Percentage of Nonoverlapping Data.

Table 3 Conditional power, averaged over 100 replications for data sets generated from a uniform distribution

Data: Uniform distribution

Design		CRD			RBD			RRAD		
N	TE	Effect Size								
		MD	NAP	PND	MD	NAP	PND	MD	NAP	PND
12	0	4.95	3.97	1.04	3.23	3.27	0.14	6.19	4.75	1.82
	0.5	10.97	9.23	3.62	5.98	6.99	0.70	10.62	11.07	6.81
	1	31.71	24.66	10.70	16.38	16.73	2.94	32.47	28.73	19.80
	1.5	60.72	51.32	28.94	38.39	39.36	9.09	61.84	50.89	38.18
	2	87.26	80.75	50.29	59.44	54.82	21.42	84.90	80.21	58.10
	2.5	98.57	95.47	77.37	79.68	81.21	46.91	97.09	93.73	85.56
20	0	5.00	4.51	1.60	4.86	4.47	0.59	6.01	5.04	2.23
	0.5	17.26	15.60	8.49	15.48	14.52	4.05	17.40	15.25	12.02
	1	55.29	47.56	34.55	50.40	42.87	20.32	53.48	47.47	35.32
	1.5	88.51	81.38	65.31	85.89	77.38	46.08	88.28	81.44	67.37
	2	99.33	97.47	89.67	97.50	94.65	78.46	99.38	98.07	90.92
	2.5	100	99.91	98.49	99.95	99.29	96.85	100	99.90	98.82
30	0	5.13	4.87	1.88	4.99	4.62	0.89	7.59	7.44	16.23
	0.5	25.17	22.81	17.29	23.33	21.88	11.66	27.35	29.52	48.37
	1	73.54	67.53	60.74	72.39	64.36	49.13	77.24	68.14	75.43
	1.5	97.87	94.77	89.42	97.26	94.08	79.99	97.67	94.16	92.31
	2	99.99	99.83	98.39	99.99	99.64	97.54	99.97	99.60	98.14
	2.5	100.0	100	99.98	100	100	99.87	100	100	99.89
40	0	5.03	4.96	1.79	5.00	4.77	1.05	5.44	4.87	2.40
	0.5	31.80	30.50	28.84	30.98	29.12	18.16	32.96	30.61	32.99
	1	86.25	81.28	81.04	84.59	78.46	70.28	86.90	78.87	83.09
	1.5	99.76	98.57	98.50	99.57	98.59	97.53	99.76	98.37	98.31
	2	100	99.99	99.96	100	99.94	99.87	100	100	99.96
	2.5	100	100	100	100	100	100	100	100	100

N: number of measurement occasions; TE: Treatment Effect; CRD: Completely Randomized Design; RBD: Randomized Block Design; RRAD: Restricted Randomized Alternation Design; MD: Mean Difference; NAP: Nonoverlap of All Pairs; PND: Percentage of Nonoverlapping Data.

Table 4 Conditional power, averaged over 100 replications for data sets from an AR1 model with AR = 0.6

Data: AR1 model with AR = 0.6

Design		CRD			RBD			RRAD		
N	TE	Effect Size								
		MD	NAP	PND	MD	NAP	PND	MD	NAP	PND
12	0	4.79	4.07	1.10	3.29	3.18	0.02	4.56	4.27	2.22
	0.5	12.26	10.11	4.87	9.31	11.06	0.29	14.39	12.82	8.20
	1	31.34	25.24	13.82	33.51	25.01	4.64	38.55	31.59	20.26
	1.5	57.69	55.50	33.59	55.33	54.98	8.38	67.02	65.08	41.82
	2	80.37	75.68	50.58	79.54	84.16	21.88	87.81	82.93	61.86
	2.5	92.85	89.59	67.83	89.17	93.07	45.11	95.29	93.88	74.04
20	0	5.12	4.54	1.60	4.98	4.30	0.23	4.66	4.88	2.61
	0.5	16.33	14.50	8.84	24.87	19.75	1.76	20.70	19.62	9.53
	1	48.28	41.68	19.82	70.40	61.91	9.12	59.79	59.00	27.40
	1.5	79.32	78.09	48.60	97.12	94.16	32.73	89.90	88.07	54.92
	2	94.67	93.57	71.05	99.91	99.86	59.13	98.75	98.69	81.32
	2.5	98.50	97.87	86.41	100	99.99	80.20	99.93	99.83	89.86
30	0	5.10	4.75	1.93	4.98	4.56	0.27	6.83	6.68	16.78
	0.5	20.26	19.82	11.45	35.86	30.33	2.91	36.13	31.97	31.75
	1	60.87	58.58	29.45	90.27	82.59	18.48	83.19	79.47	52.38
	1.5	90.47	89.53	55.43	99.89	99.33	46.73	98.54	98.01	74.02
	2	98.84	97.92	78.54	100	100	69.55	99.95	99.88	88.58
	2.5	99.88	99.91	91.96	100	100	90.80	100	100	93.76
40	0	4.90	4.85	1.70	4.99	4.71	0.27	4.08	4.45	2.18
	0.5	25.52	25.47	10.72	48.34	45.39	5.96	38.22	33.64	11.64
	1	71.50	68.41	32.84	95.28	95.31	21.57	91.46	89.13	37.69
	1.5	96.21	94.67	59.26	100	99.96	56.04	99.89	99.79	64.76
	2	99.78	99.65	83.73	100	100	80.90	100	100	89.37
	2.5	100	99.99	91.55	100	100	91.21	100	100	95.92

N: number of measurement occasions; TE: Treatment Effect; CRD: Completely Randomized Design; RBD: Randomized Block Design; RRAD: Restricted Randomized Alternation Design; MD: Mean Difference; NAP: Nonoverlap of All Pairs; PND: Percentage of Nonoverlapping Data.

References

Adams, D. C., & Anthony, C. D. (1996). Using randomization techniques to analyse behavioural data. *Animal Behaviour, 51,* 733–738.

Agresti, A. (1992). A survey of exact inference for contingency tables. *Statistical Science, 7,* 131–153.

Allison, D. B., & Gorman, B. S. (1993). Calculating effect sizes for meta-analysis: The case of the single case. *Behaviour Research Therapy, 31,* 621–631.

Alnahdi, G. H. (2015). Single-subject design in special education: Advantages and limitations. *Journal of Research in Special Educational Needs, 15,* 257–265.

Barlow, D. H., Nock, M. K., & Hersen, M. (2009). *Single case experimental designs: Strategies for studying behavior change* (3rd ed.). Boston: Pearson.

Borckardt, J. J., & Nash, M. R. (2014). Simulation modelling analysis for small sets of single-subject data collected over time. *Neuropsychological Rehabilitation, 24,* 492–506.

Bowman-Perrott, L., Burke, M. D., de Marin, S., Zhang, N., & Davis, H. (2015). A meta-analysis of single-case research on behavior contracts: Effects on behavioral and academic outcomes among children and youth. *Behavior Modification, 39,* 247–269. doi:10.1177/0145445514551383

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

Busk, P. L., & Serlin, R. C. (1992). Meta-analysis for single-case research. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 187–212). HillsdaleJ: Erlbaum.

Campbell, J. M. (2013). Commentary on PND at 25. *Remedial and Special Education, 34,* 20–25.

Center, B. A., Skiba, R. J., & Casey, A. (1985–1986). A methodology for the quantitative synthesis of intra-subject design research. *Journal of Special Education, 19,* 387–400.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale: Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112,* 155–159. doi:10.1037/0033-2909.112.1.155

Collings, B. J., & Hamilton, M. A. (1988). Estimating the power of the two-sample Wilcoxon test for location shift. *Biometrics, 44,* 847–860.

Corcoran, C. D., & Mehta, C. R. (2002). Exact level and power of permutation, bootstrap, and asymptotic tests of trend. *Journal of Modern Applied Statistical Methods*, *1*(1). Retrieved from digitalcommons.wayne.edu/jmasm/vol1/iss1/7

Cox, D. R., & Reid, N. (2000). *The theory of the design of experiments*. Boca Raton: Chapman & Hall/CRC.

Dugard, P. (2014). Randomization tests: A new gold standard? *Journal of Contextual Behavioral Science, 3,* 65–68.

Edgington, E. S. (1967). Statistical inference from $N = 1$ experiments. *Journal of Psychology, 65,* 195–199.

Edgington, E. S. (1969). Approximate randomization tests. *Journal of Psychology, 72,* 143–149.

Edgington, E. S., & Onghena, P. (2007). *Randomization tests* (4th ed.). Boca Raton: Chapman & Hall/CRC.

Fabiano, G. A., Pelham, W. E., Coles, E. K., Gnagy, E. M., Chronis-Tuscano, A., & O'Connor, B. C. (2009). A meta-analysis of behavioral treatments for attention-deficit/hyperactivity disorder. *Clinical Psychology Review, 29,* 129–140.

Ferron, J. M., Farmer, J. L., & Owens, C. M. (2010). Estimating individual treatment effects from multiple-baseline data: A Monte Carlo study of multilevel-modeling approaches. *Behavior Research Methods, 42,* 930–943. doi:10.3758/BRM.42.3.930

Ferron, J. M., & Levin, J. R. (2014). Single-case permutation and randomization statistical tests: Present status, promising new developments. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: Methodological and statistical advances* (pp. 153–183). Washington, DC: American Psychological Association.

Ferron, J., & Onghena, P. (1996). The power of randomization tests for single-case phase designs. *Journal of Experimental Education, 64,* 231–239.

Ferron, J., & Sentovich, C. (2002). Statistical power of randomization tests used with multiple-baseline designs. *Journal of Experimental Education, 70,* 165–178.

Ferron, J., & Ware, W. (1995). Analyzing single-case data: The power of randomization tests. *Journal of Experimental Education, 63,* 167–178.

Gabler, N. B., Duan, N., Vohra, S., & Kravitz, R. L. (2011). *N*-of-1 trials in the medical literature: A systematic review. *Medical Care, 49,* 761–768. doi:10.1097/MLR.0b013e318215d90d

Gabriel, K. R., & Hsu, C.-F. (1983). Evaluation of the power of rerandomization tests, with application to weather modification experiments. *Journal of the American Statistical Association, 78,* 766–775.

Gast, D. L., & Ledford, J. R. (2014). *Single case research methodology: Applications in special education and behavioral sciences* (2nd ed.). New York: Routledge.

Gottman, J. M., & Glass, G. V. (1978). Analysis of interrupted time-series experiments. In T. R. Kratochwill (Ed.), *Single-subject research: Strategies for evaluating change* (pp. 197–237). New York: Academic Press.

Hamilton, M. A., & Collings, B. J. (1991). Determining the appropriate sample size for nonparametric tests for location shift. *Technometrics, 3,* 327–337.

Hammond, D., & Gast, D. L. (2010). Descriptive analysis of single-subject research designs: 1983–2007. *Education and Training in Autism and Developmental Disabilities, 45,* 187–202.

Hedges, L. V., Pustejovsky, J. E., & Shadish, W. R. (2012). A standardized mean difference effect size for single case designs. *Research Synthesis Methods, 3,* 324–239.

Heyvaert, M., Maes, B., Van den Noortgate, W., Kuppens, S., & Onghena, P. (2012). A multilevel meta-analysis of single-case and small-*n* research on interventions for reducing challenging behavior in persons with intellectual disabilities. *Research in Developmental Disabilities, 33,* 766–780.

Heyvaert, M., Moeyaert, M., Verkempynck, P., Van den Noortgate, W., Vervloet, M., Ugille, M., & Onghena, P. (2017). Testing the intervention effect in single-case experiments: A Monte Carlo simulation study. *Journal of Experimental Education, 85,* 175–196. doi:10.1080/00220973.2015.1123667

Heyvaert, M., & Onghena, P. (2014). Analysis of single-case data: Randomisation tests for measures of effect size. *Neuropsychological Rehabilitation, 24,* 507–527.

Heyvaert, M., Saenen, L., Campbell, J. M., Maes, B., & Onghena, P. (2014). Efficacy of behavioral interventions for reducing problem behavior in persons with autism: An updated quantitative synthesis of single-subject research. *Research in Developmental Disabilities, 35,* 2463–2476.

Heyvaert, M., Saenen, L., Maes, B., & Onghena, P. (2014). Systematic review of restraint interventions for challenging behaviour among persons with intellectual disabilities: Focus on effectiveness in single-case experiments. *Journal of Applied Research in Intellectual Disabilities, 27,* 493–510.

Heyvaert, M., Wendt, O., Van den Noortgate, W., & Onghena, P. (2015). Randomization and data-analysis items in quality standards for single-case experimental studies. *Journal of Special Education, 49,* 146–156.

Hinkelmann, K., & Kempthorne, O. (2008). *Design and analysis of experiments: Vols. I and II* (2nd ed.). Hoboken: Wiley.

Horner, R. H., Swaminathan, H., Sugai, G., & Smolkowski, K. (2012). Considerations for the systematic analysis and use of single-case research. *Education and Treatment of Children, 35,* 269–290.

Kazdin, A. E. (2011). *Single-case research designs: Methods for clinical and applied settings* (2nd ed.). New York: Oxford University Press.

Keller, B. (2012). Detecting treatment effects with small samples: The power of some tests under the randomization model. *Psychometrika, 2,* 324–338.

Kempthorne, O. (1955). The randomization theory of experimental inference. *Journal of the American Statistical Association, 50,* 946–967.

Kempthorne, O., & Doerfler, T. E. (1969). The behavior of some significance tests under experimental randomization. *Biometrika, 56,* 231–248.

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 the What Works Clearinghouse website: ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf

Kratochwill, T. R., & Levin, J. R. (1992). *Single-case research design and analysis: New directions for psychology and education*. Hillsdale: Erlbaum.

Krauth, J. (1988). *Distribution-free statistics: An application-oriented approach*. New York: Elsevier.

Lachin, J. M. (2005). A review of methods for futility stopping based on conditional power. *Statistics in Medicine, 24,* 2747–2764.

Lehmann, E. L. (1959). *Testing statistical hypotheses*. Hoboken: Wiley.

Lenz, A. (2012). Calculating effect size in single-case research: A comparison of nonoverlap methods. *Measurement and Evaluation in Counseling and Development, 46,* 64–73.

Leong, H. M., Carter, M., & Stephenson, J. (2015). Systematic review of sensory integration therapy for individuals with disabilities: Single case design studies. *Research in Developmental Disabilities, 47,* 334–351.

Levin, J. R., Ferron, J. M., & Gafurov, B. S. (2014). Improved randomization tests for a class of single-case intervention designs. *Journal of Modern Applied Statistical Methods, 13,* 2–52.

Levin, J. R., Ferron, J. M., & Kratochwill, T. R. (2012). Nonparametric statistical tests for single-case systematic and randomized ABAB … AB and alternating treatment intervention designs: New developments, new directions. *Journal of School Psychology, 50,* 599–624.

Maggin, D. M., O'Keeffe, B. V., & Johnson, A. H. (2011). A quantitative synthesis of methodology in the meta-analysis of single-subject research for students with disabilities: 1985–2009. *Exceptionality, 19,* 109–135.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics, 18,* 50–60.

Manolov, R., Solanas, A., Bulté, I., & Onghena, P. (2010). Data-division-specific robustness and power of randomization tests for ABAB designs. *Journal of Experimental Education, 78,* 191–214.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105,* 156–166. doi:10.1037/0033-2909.105.1.156

Michiels, B., Heyvaert, M., Meulders, A., & Onghena, P. (2017). Confidence intervals for single-case effect size measures based on randomization test inversion. *Behavior Research Methods, 49,* 363–381. doi:10.3758/s13428-016-0714-4

Moeller, J. D., Dattilo, J., & Rusch, F. (2015). Applying quality indicators to single-case research designs used in special education: A systematic review. *Psychology in the Schools, 52,* 139–153.

Onghena, P. (1992). Randomization tests for extensions and variations of ABAB single-case experimental designs: A rejoinder. *Behavioral Assessment, 14,* 153–171.

Onghena, P. (1994). *The power of randomization tests for single-case designs (Unpublished doctoral dissertation)*. Leuven: Catholic University of Leuven.

Onghena, P. (2005). Single-case designs. In B. Everitt & D. Howell (Eds.), *Encyclopedia of statistics in behavioral science : vol. 4* (pp. 1850–1854). Chichester: Wiley.

Onghena, P., & Edgington, E. S. (1994). Randomization tests for restricted alternating treatments designs. *Behaviour Research and Therapy, 32,* 783–786.

Onghena, P., & Edgington, E. S. (2005). Customization of pain treatments: Single-case design and analysis. *Clinical Journal of Pain, 21,* 56–68.

Parker, R. I., Hagan-Burke, S., & Vannest, K. (2007). Percent of all non-overlapping data (PAND): An alternative to PND. *Journal of Special Education, 40,* 194–204.

Parker, R. I., & Vannest, K. J. (2009). An improved effect size for single-case research: Nonoverlap of all pairs. *Behavior Therapy, 40,* 357–367.

Parker, R. I., Vannest, K. J., Davis, J. L., & Sauber, S. B. (2011). Combining nonoverlap and trend for single-case research: Tau-U. *Behavior Therapy, 42,* 284–299.

Parker, R. I., Vannest, K. J., & Brown, L. (2009). The improvement rate difference for single case research. *Exceptional Children, 75,* 135–150.

Pesarin, F., & De Martini, D. (2002). On unbiasedness and power of permutation tests. *Metron, 60,* 3–19.

Pratt, J. W., & Gibbons, J. D. (1981). *Concepts of nonparametric theory*. New York: Springer.

Schlosser, R. W., Lee, D. L., & Wendt, O. (2008). Application of the percentage of non-overlapping data (PND) in systematic reviews and meta-analyses: A systematic review of reporting characteristics. *Evidence-Based Communication Assessment and Intervention, 2,* 163–187.

Scruggs, T. E., Mastropieri, M. A., & Casto, G. (1987). The quantitative synthesis of single subject research: Methodology and validation. *Remedial and Special Education, 8,* 24–33.

Senchaudhuri, P., Mehta, C. R., & Patel, N. R. (1995). Estimating exact p-values by the method of control variates, or Monte Carlo rescue. *Journal of the American Statistical Association, 90,* 640–648.

Shadish, W. R. (2014). Analysis and meta-analysis of single-case designs: An introduction. *Journal of School Psychology, 52,* 109–122.

Shadish, W. R., Rindskopf, D. M., & Hedges, L. V. (2008). The state of the science in the meta-analysis of single-case experimental designs. *Evidence-Based Communication Assessment and Intervention, 2,* 188–196.

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

Shamseer, L., Sampson, M., Bukutu, C., Schmid, C. H., Nikles, J., Tate, R., & CENT Group. (2016). CONSORT extension for reporting *N*-of-1 trials (CENT) 2015: Explanation and elaboration. *Journal of Clinical Epidemiology, 76,* 18–46. doi:10.1016/j.jclinepi.2015.05.018

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

Solanas, A., Manolov, R., & Onghena, P. (2010). Estimating slope and level change in $N = 1$ designs. *Behavior Modification, 34,* 195–218.

Solomon, B. G. (2014). Violations of assumptions in school-based single-case data: Implications for the selection and interpretation of effect sizes. *Behavior Modification, 38,* 477–496. doi:10.1177/0145445513510931

Swaminathan, H., & Rogers, H. J. (2007). Statistical reform in school psychology research: A synthesis. *Psychology in the Schools, 44,* 543–549.

Tate, R., Togher, L., Perdices, M., McDonald, S., & Rosenkoetter, U. (2012). *Developing reporting guidelines for single-case experimental designs: The SCRIBE project*. Paper presented at the 9th Conference of the Neuropsychological Rehabilitation Special Interest Group of the World Federation for Neurorehabilitation, Bergen, Norway

Van den Noortgate, W., & Onghena, P. (2003). Hierarchical linear models for the quantitative integration of effect sizes in single-case research. *Behavior Research Methods, Instruments, & Computers, 35,* 1–10. doi:10.3758/BF03195492

Welch, W., & Gutierrez, L. G. (1988). Robust permutation tests for matched-pairs designs. *Journal of the American Statistical Association, 402,* 450–455.

White, D. M., Rusch, F. R., Kazdin, A. E., & Hartmann, D. P. (1989). Applications of meta-analysis in individual-subject research. *Behavioral Assessment, 11,* 281–296.

Wolery, M., Busick, M., Reichow, R., & Barton, E. E. (2010). Comparison of overlap methods for quantitatively synthesizing single-subject data. *Journal of Special Education, 44,* 18–28.