Handling missing data in randomization tests for single-case experiments: A simulation study

Tamal Kumar De¹ · Bart Michiels¹ · René Tanious¹ · Patrick Onghena¹

Published online: 2 January 2020
© The Psychonomic Society, Inc. 2020

Abstract

Single-case experiments have become increasingly popular in psychological and educational research. However, the analysis of single-case data is often complicated by the frequent occurrence of missing or incomplete data. If missingness or incompleteness cannot be avoided, it becomes important to know which strategies are optimal, because the presence of missing data or inadequate data handling strategies may lead to experiments no longer “meeting standards” set by, for example, the What Works Clearinghouse. For the examination and comparison of strategies to handle missing data, we simulated complete datasets for ABAB phase designs, randomized block designs, and multiple-baseline designs. We introduced different levels of missingness in the simulated datasets by randomly deleting 10%, 30%, and 50% of the data. We evaluated the type I error rate and statistical power of a randomization test for the null hypothesis that there was no treatment effect under these different levels of missingness, using different strategies for handling missing data: (1) randomizing a missing-data marker and calculating all reference statistics only for the available data points, (2) estimating the missing data points by single imputation using the state space representation of a time series model, and (3) multiple imputation based on regressing the available data points on preceding and succeeding data points. The results are conclusive for the conditions simulated: The randomized-marker method outperforms the other two methods in terms of statistical power in a randomization test, while keeping the type I error rate under control.

Keywords Single-case data · Randomization test · Missing data · Simulation study · Power analysis

Single-case experiments

A single-case experiment (SCE) is an experiment in which the effect of applying different levels (treatments) of at least one independent variable is observed repeatedly on a single entity over a period of time (Barlow, Nock, & Hersen, 2009; Kazdin, 2011; Kratochwill & Levin, 1992; Onghena, 2005). The

purpose of such an experiment is to investigate a causal relationship between the manipulated variable(s) and one or more repeatedly observed outcome variables. SCEs are often conducted in the behavioral sciences, particularly in subdisciplines such as clinical psychology, counseling psychology, neuropsychology, school psychology, psychopharmacology, social work, education, and special education (Heyvaert & Onghena, 2014). Heyvaert and Onghena also pointed out that the low cost of an SCE, its feasibility for studying rare conditions and within-subjects variability, and the present-day availability of data analysis techniques and computer programs have led to a steady increase in interest in SCE research in recent years.

A plethora of methods for SCE data analysis are in vogue nowadays. These methods can be categorized in two broad groups: visual methods and statistical methods (Heyvaert, Wendt, Van den Noortgate, & Onghena, 2015). Visual methods refer to the assessment of level, trend, variability, and other aspects of the data to assess the existence and consistency of treatment effects (Heyvaert et al., 2015; Horner, Swaminathan, Sugai, & Smolkowski, 2012; Kratochwill

✉ Tamal Kumar De
tamalkumar.de@kuleuven.be

Bart Michiels
bart.michiels@kuleuven.be

René Tanious
rene.tanious@kuleuven.be

Patrick Onghena
patrick.onghena@kuleuven.be

¹ Methodology of Educational Sciences Research Group, Faculty of Psychology and Educational Sciences, KU Leuven, Leuven, Belgium

et al., 2010). Traditionally, applied SCE researchers have relied on visual analysis to establish the existence and magnitude of causal relationships (Kratochwill et al., 2010). However, visual analysis is subjective, often unreliable and is prone to overestimating the effect (Smith, 2012). Hence, rigorous and objective statistical analysis methods are pertinent for SCEs.

Statistical methods for SCEs can be broadly categorized into three categories: effect size calculation, statistical modeling, and statistical inference (Michiels, Heyvaert, & Onghena, 2018). Effect size measures for SCEs include mean-difference-based measures such as Cohen's d and Hedges's g (Heyvaert et al., 2015), nonoverlap-based measures such as percentage of nonoverlapping data (Scruggs, Mastropieri, & Casto, 1987), percentage exceeding the median (Ma, 2006), and nonoverlap of all pairs (Parker & Vannest, 2009). Methods for statistical modeling of SCEs include multilevel modeling, structural equation modeling, and interrupted time series analysis (Heyvaert et al., 2015; Michiels et al., 2018; Van den Noortgate & Onghena, 2003). Finally, statistical inference methods include methods for conducting statistical significance tests and calculating confidence intervals. Traditional group comparison methods like t and F tests have been used for SCEs; however, due to the serial dependency in SCE data (Busk & Marascuilo, 1988; Solomon, 2014) and plausible violations of the distributional assumptions of the traditional tests in behavioral experiments (Adams & Anthony, 1996; Smith, 2012), these methods are usually not recommended (Houle, 2009). Nonparametric tests (e.g., Kruskal–Wallis tests, Wilcoxon–Mann–Whitney tests, randomization tests, and others), which by definition do not require distributional assumptions, are more suited to analyze SCE data (Heyvaert et al., 2015).

Randomization tests

In this study, we focus on a particular type of nonparametric test: the randomization test. A randomization test (RT) is a type of statistical significance test based on the random assignment of the experimental units to the levels of an independent variable (treatments) to test hypotheses about causal effects (Edgington & Onghena, 2007). In the context of SCEs, the experimental units are measurement occasions, which are randomly assigned to treatment levels. RTs obtain their statistical validity from this random assignment. A benefit of RTs is their flexibility of customization to the particular research purpose. For example, any randomization scheme can be taken into account and any test statistic can be used to test the null hypothesis that the manipulated variable does not causally affect the observed outcome variable (Edgington & Onghena, 2007).

We refer the interested reader to the monograph of Edgington and Onghena (2007) and to the tutorial of Heyvaert and Onghena (2014) for comprehensive overviews of the RT procedure and its properties. For the present article, it suffices to recall the four steps involved in an RT for an SCE. (1) First, all possible random assignments of measurement occasions to treatment levels are recorded and an appropriate test statistic is chosen. (2) Next, one assignment out of the possible assignments is chosen at random and the experiment is conducted with that assignment. (3) Once the data have been collected, the chosen test statistic is calculated for each of the possible assignments. (4) Finally, the p value is calculated as the percentage of all possible test statistic values that are at least as extreme as the observed test statistic.

Missing data in SCEs

Missing data are common in SCEs, often due to human factors such as forgetfulness or systematic noncompliance to the protocols of data collection in an experiment (Smith, 2012). In their review of published SCE research, Peng and Chen (2018) found that 24% (34 out of 139) of articles reported missing data, and another 7% (10 out of 139) did not provide sufficient information to determine whether missing data were present. In a review article regarding missing data in educational research, although it was not restricted to SCEs, Peugh and Enders (2004) reported that out of 989 published articles in 1999, 160 (16.18%) studies had missing data, and 92 (9.3%) were indeterminate. They also reported that among the studies that did report missing data, the percentage of missing cases (i.e., the percentage of cases containing missing data, not the percentage of missing measurements) ranged from 1% to 67%, with a mean percentage of 7.60%.

Failing to account for missing data or handling missing data improperly can lead to biased, inefficient, or unreliable results (Little & Rubin, 2002; Schafer & Graham, 2002). Improper missing-data handling techniques can further weaken the generalizability of study results (Peng & Chen, 2018; Rubin, 1987). Smith (2012) specifically mentioned missing-data methods prevalent in SCE research and advocated the use of the maximum likelihood and expectation maximization (EM) methods. Smith, Borckardt, and Nash (2012) conducted a simulation study and concluded that the EM algorithm is effective at replacing missing data, while maintaining statistical power in autocorrelated SCE data streams. Chen, Feng, Wu, and Peng (2019) extended the results of Smith et al. (2012), exploring the relative bias and root-mean-squared error of the estimates and the relative bias of the estimated standard errors using the EM algorithm, and concluded that the EM algorithm performs adequately when

applied to AB phase design data. Peng and Chen demonstrated the use of multiple imputation (MI) for handling missing data in a published single-case ABAB phase design and discussed several advantages of this method: MI retains the design structure of the study and collected data, avoids potential bias, and captures uncertainty concerning the imputed scores.

Missing-data handling methods

In this study, we evaluated the performance of different missing-data handling methods for SCEs analyzed by RTs in a simulation study. We decided to test methods that are relevant to RTs. Additionally, we only considered methods that either are simple to implement or have existing implementations in code available for everyone to use, preferably in the statistical computing language and environment R. Three methods were selected under these conditions: the randomized-marker method, single imputation using an autoregressive integrated moving average (ARIMA) model, and MI using multivariate imputation by chained equations (MICE).

The randomized-marker method A simple method was proposed by Edgington and Onghena (2007). In this method, a regular randomization test is performed, but instead of randomly reshuffling numerical scores only, the missing data are included by representing them with a certain code or marker (e.g., “NA” for “not available”). The method reshuffles the markers together with the numerical scores and performs all calculations taking into account the missing data in one or more of the treatment conditions. This procedure leads to valid p values because the randomization test null hypothesis of no differential effect of the treatments applies to both the numerical scores and the missingness. Consequently, under the null hypothesis, the missingness of a data point is independent of the assignment to a particular treatment at that particular time point. The value of the test statistic for a particular reshuffle may be calculated, while taking into account the random division of the missing data among the different treatments. This method has been implemented in the SCRT R package (Bulté & Onghena, 2008). We will refer to this method as the “randomized-marker method.”

We demonstrate this method using a hypothetical dataset from an SCE following a randomized block design (RBD) for the comparison of two treatments, namely A and B. In an RBD with two treatments, those treatments are paired among successive measurements, with random determination of the order of the treatments in the pair (Onghena, 2005). In an experiment with six measurements, treatments A and B can be administered in three

Table 1 Hypothetical randomized block design experiment with six measurement occasions and eight possible assignments

Time	Data	R1	R2	R3	R4	R5	R6	R7	R8
1	5	A	B	A	B	A	B	A	B
2	3	B	A	B	A	B	A	B	A
3	NA	A	A	B	B	A	A	B	B
4	2	B	B	A	A	B	B	A	A
5	6	A	A	A	A	B	B	B	B
6	NA	B	B	B	B	A	A	A	A
	A – B	3	1	1.33	– 1.33	1.33	– 1.33	– 1	– 3

blocks, resulting in $2^3 = 8$ possible randomizations of the treatments (see Table 1). Let us assume the experiment was conducted with the randomization A, B, A, B, A, B (R1 in Table 1), and the observed scores were 5, 3, 2, and 6 for measurement occasions 1, 2, 4, and 5, respectively, and missing for measurement occasions 3 and 6. We will use NA as the missing-data marker in the dataset. For the randomized-marker method, the NAs are also randomized along with the observed scores. So, instead of randomizing only 5, 3, 2, and 6, we randomize 5, 3, NA, 2, 6, and NA for an RT. We can now evaluate the p value for the RT with the null hypothesis that there is no treatment effect and the alternative hypothesis that treatment B leads to lower observed values than treatment A using a mean difference test statistic, $\text{mean}(A) - \text{mean}(B)$. For the randomization R1:

$$\begin{aligned}\text{mean}(A) - \text{mean}(B) &= \text{mean}(5, \text{NA}, 6) - \text{mean}(3, 2, \text{NA}) \\ &= \text{mean}(5, 6) - \text{mean}(3, 2) = 3.\end{aligned}$$

The test statistic values for the other randomizations can be calculated in the same way (see the last row of Table 1). The p value can be calculated as the percentage of test statistic values that are at least as large as the observed test statistic for R1. Since R1 is the only randomization that leads to a test statistic as large as the observed test statistic (also R1), out of eight possible randomizations, the p value is $1/8 = .125$. The null hypothesis in this experiment cannot be rejected using the standard level of significance of .05. In fact, since .125 is the lowest possible p value for this experiment, we can conclude that this experiment lacks power to reject the null hypothesis at significance levels of .01, .05, and even .1.

Single imputation using an ARIMA model Single imputation (SI) is one of the most frequently used methods of handling missing data. Last observation carried forward, mean imputation, hot deck imputation, and regression

imputation are examples of this method (Roth, 1994; Schafer & Graham, 2002). Because the data from SCEs often demonstrate high levels of autocorrelation (Busk & Marascuilo, 1988; Heyvaert & Onghena, 2014; Solomon, 2014), we decided to implement a time series (TS)-based method for imputation. Harvey and Pierse (1984) proposed a two-step procedure: fitting an ARIMA model to the available data, and applying Kalman smoothing to the state space representation of the fitted ARIMA model in order to estimate the missing data. The ARIMA model is one of the more popular TS models (Houle, 2009; Shumway & Stoffer, 2017), and it combines autoregressive, moving average, and nonstationary models into a single model (Box & Jenkins, 1970). Due to the autoregressive component, an ARIMA model is well suited to modeling autocorrelated data. Additionally, most statistical software packages provide functions for estimating the parameters in an ARIMA model.

An autoregressive moving average (ARMA) model with order (p, q) of a time series x_t can be represented by the following equation,

$$x_t = \Phi_1 x_{t-1} + \dots + \Phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + \dots + \theta_q w_{t-q}, (1)$$

where w_i are Gaussian noise variables that follow $N(0, \sigma^2)$, whereas Φ_i and θ_i are constants. An ARIMA model with order (p, d, q) , which incorporates differencing, can be written as

$$\Phi(B)(1-B)^d x_t = \theta(B)w_t, (2)$$

where B represents the backshift operator, such that $Bx_t = x_{t-1}$, $\Phi(B) = 1 - \Phi_1 B - \dots - \Phi_p B^p$, and $\theta(B) = 1 + \theta_1 B + \dots + \theta_q B^q$. A simple state space representation of the ARIMA model, as was proposed by Box and Jenkins (1970), can be denoted as $X_t = [1, 0, \dots, 0]Y_t$. Here, Y_t is defined as

$$Y_t = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \dots & 0 & 1 \\ \varphi_r & \varphi_{r-1} & \dots & \dots & \varphi_1 \end{bmatrix} Y_{t-1} + \begin{bmatrix} 1 \\ \psi_{r-1} \\ \vdots \\ \psi_1 \end{bmatrix} W_t, (3)$$

where $r = \max(p + d, q + 1)$ and φ_i and ψ_i are constant coefficients. The model coefficients can be estimated using a Kalman filter (Kalman, 1960) algorithm, developed for ARIMA models in Gardner, Harvey, and Phillips (1980). Finally, a smoothing algorithm (Anderson & Moore, 1979) can be applied in order to impute the missing values. A detailed presentation of these algorithms is beyond the scope of the present article, but the interested reader is referred to Harvey and Pierse (1984) for details

on this method. We utilized the available implementation of this method in R by Moritz and Bartz-Beielstein (2017).

MI using MICE MI was proposed by Rubin (1987) and is a Monte Carlo technique in which missing values are replaced by multiple simulated values in order to create multiple versions of the complete dataset; each of the simulated versions is analyzed separately and then combined to generate estimates that incorporate uncertainty due to missingness (Schafer, 1999). When the data series is univariate, it is possible to use MI by sampling imputations from observed data. MI is also suitable for multivariate datasets, because data on covariates can be used to predict the missing data points. Traditionally, there are two methods for multivariate data: joint modeling and fully conditional specification, also known as MICE. We decided to use the MICE implementation developed by van Buuren and Groothuis-Oudshoorn (2011), because this method is easy to understand and the implementation in R is already freely available. This method works by using covariate data to impute the missing data. However, the simulated data in our study are univariate. Hence, we decided to add a leading and a lagging vector, consisting of subsequent and previous scores, to the dataset as covariates, to exploit the previously discussed autoregressive property of SCE data in order to impute the missing values. In our example dataset (Table 1), this would mean that for the observed scores (5, 3, NA, 2, 6, NA), the two covariate vectors would be (NA, 5, 3, NA, 2, 6) and (3, NA, 2, 6, NA, NA). The missing data in each vector are imputed by linearly regressing this vector on the remaining vectors, repeating this process for the other vectors, and then iterating this entire process again and again. The iterative regressions are kicked off using random samples to fill in the missing data in the first iteration.

The guidelines offered by Rubin (1987) and further discussed by Schafer (1999) suggest that five to ten imputations are sufficient. Contrary to that approach, recent research has suggested that a larger number of imputations might be beneficial (Graham, Olchowski, & Gilreath, 2007). However, Graham et al. also estimated that, at 50% missing data, the difference between the power calculated using ten imputations and the power calculated using 100 imputations is only around 5%. Hence, taking into account the linear increase in computation time with an increase in the number of imputations, we chose to simulate ten imputations of the missing data. The final test statistic was calculated by taking the average of ten test statistics, one from each of the imputed datasets. The reference distribution was also constructed in the same way.

Objectives and research questions

Our objective in this study was to compare three methods of handling missing data in RTs for SCEs. The simplest metrics by which we can measure the performance of any hypothesis test, such as an RT, is by estimating its statistical power and type I error probability. We wanted to find out which missing-data method resulted in the highest values for power under similar simulation conditions, while keeping type I errors under control. Additionally, since RTs are computationally intensive, we were interested in finding out which method resulted in the least increase in computation time. Hence, we also recorded the run times of each of the simulation conditions as a secondary performance metric.

Method

Power and type I error

The power of a statistical hypothesis test is defined as the probability that a false null hypothesis is rejected. The probability of a type I error is defined as the probability that a true null hypothesis is rejected. Multiple studies have evaluated the power of RTs in the context of SCEs without missing data (Ferron & Onghena, 1996; Ferron & Sentovich, 2002; Ferron & Ware, 1995; Michiels et al., 2018). In the present study, we followed the methodology used by Ferron and Onghena to calculate power estimates and type I error rates.

Simulation conditions

Missing-data percentage As we noted previously, Peugh and Enders (2004) reported that in 989 published articles on educational research in 1999, the percentage of cases with missing data ranged from 1% to 67%, with a mean percentage of 7.60%. In keeping with the missing-data percentages reported in such studies, the percentage of missing data was varied in three steps: 10%, 30%, and 50%, to represent low, moderate, and high levels of missing data, respectively. The missing data were created by deleting randomly from the complete dataset, so that the missing data were missing completely at random (Little & Rubin, 2002).

SCE design type Three different randomized single-case design types were chosen: an RBD with a block size of two, an ABAB phase design with randomized phase change points, and a multiple-baseline design (MBD) with four subjects (for more details on these designs for

SCEs, we refer readers to Barlow et al., 2009; Edgington & Onghena, 2007; Hammond & Gast, 2010).

The RBD, discussed previously, is most similar to the designs used for medical *N*-of-1 randomized controlled trials (Guyatt, Jaeschke, & McGinn, 2002; Guyatt et al., 1990); hence, it was included. Any alternation design with at least five repetitions of the alternating sequence would meet the standards set by Kratochwill et al. (2010), which is achieved in an RBD with a block size of two and a minimum of ten measurements.

The ABAB phase design and the MBD are two of the most popular SCE designs in the behavioral sciences; hence, we included both in the study (Hammond & Gast, 2010; Shadish & Sullivan, 2011). In a phase design, the sequence of measurements is divided into treatment phases, with several measurements with the same treatment level being applied in each phase (Onghena & Edgington, 2005). The ABAB phase design, as the name implies, consists of four such phases: a baseline (A) phase, a treatment (B) phase, followed by another baseline (A) phase, and another treatment (B) phase. In accordance with the guidelines recommended by the What Works Clearinghouse panel of experts (WWC; Kratochwill et al., 2010), the ABAB phase designs used in the simulation study had a minimum of three measurements per phase.

The MBD is a simultaneous replication design that consists of multiple AB phase designs implemented simultaneously, whereas the intervention is started sequentially across subjects (Onghena & Edgington, 2005). Whereas the simulated datasets for both RBD and ABAB phase designs consisted of data for one subject, the number of subjects in MBD was chosen to be four, on the basis of reviews of published research; Ferron, Farmer, and Owens (2010) reported a median of four subjects in MBD experiments, and Shadish and Sullivan (2011) reported an average of 3.64 cases in SCE studies. The restriction of a minimum of three measurements per phase was also applied to MBD. As a result, any measurement occasion after the third measurement and before the third-to-last measurement could be a possible intervention start point. We randomly assigned unique intervention start points to each subject from this range of possible intervention start points. Known as the *restricted Marascuilo–Busk procedure*, this randomization procedure is recommended by Levin, Ferron, and Gafurov (2018).

Data model Three data models were included in the simulation study: a standard normal model, a uniform model with a standard deviation of 1, and an autoregressive model with order 1 and standard normal errors (AR1). The standard normal model and uniform-data model were included due to their

ubiquity as simulation models; previous studies on power using these data models could serve as benchmark for our results (Keller, 2012; Michiels et al., 2018). The uniform model also represents a simple case in which the normality assumption for observed data is grossly violated, as such violations are common among SCEs in behavioral experiments (Adams & Anthony, 1996; Smith, 2012). As has been mentioned previously, SCEs also demonstrate high levels of autocorrelation (Busk & Marascuilo, 1988; Shadish & Sullivan, 2011; Solomon, 2014). Hence, we wanted to include data models that simulate data with autocorrelation; however, we also needed to keep the study computationally manageable. Keeping these constraints in mind, we decided to include one AR1 model, which is a simple modification of the normal model generating autocorrelated data, with a relatively high autocorrelation of .6.

Number of measurements We included three different options for the number of measurements per dataset: 20, 30, and 40. Shadish and Sullivan (2011) reported a median of 20 measurements and a range of 2–160 in 809 SCEs. Interestingly, 90.6% of these SCEs had 49 or fewer measurements. Ferron et al. (2010) reported a median of 24 measurements and a range 7–58 in 75 MBD studies. We chose 20, 30, and 40 so that we could adhere to these ranges while ensuring that sufficient randomizations were possible for all design choices. Since the RBD is a balanced design, the numbers of measurements assigned to both treatment levels A and B were exactly half the number of all measurements. However, for ABAB and MBD, the numbers of measurements assigned to treatment levels A and B were decided by the chosen randomization scheme. Due to the restriction on the minimum number of measurements per phase, this number ranged from a minimum of six to a maximum of six less than the total number of measurements for the ABAB phase design, and from a minimum of three to a maximum of three less than the total number of measurements for the MBD.

Effect size Three additive effect sizes were chosen: 0, 1, and 2. We followed the unit treatment additive model for the treatment effect (Michiels et al., 2018; Welch & Gutierrez, 1988). According to this model,

$$X_i^B = X_i^A + d, \quad (4)$$

where X_i^A and X_i^B represent the expected scores if the i th measurement was assigned to the baseline (A treatment level) and to the intervention (B treatment level), respectively, and d represents the additive treatment effect. The effect size of zero was included for the purpose of calculating the type I error rate for each of the simulation conditions. The effect sizes of 1 and 2 were chosen for calculating power.

Test statistic Two test statistics were calculated: mean difference (MD) and nonoverlap of all pairs (NAP; Parker & Vannest, 2009). MD is the difference between the means of the scores in the two treatment levels, as discussed in the example in Table 1. NAP is calculated by first calculating all possible pairs of A and B treatment level scores, and then calculating the percentage of pairs for which the B scores are bigger than the A scores (or vice versa, when the A scores are expected to be bigger due to a negative effect size). These effect size measures were chosen on the basis of the recommendation by Michiels et al. (2018), who showed the superiority of these two measures over the percentage of nonoverlapping data (Scruggs et al., 1987) in terms of power in an RT. For MBD, a combined test statistic was calculated as the mean of the test statistics from each subject. Both test statistics were calculated to be symmetric. This meant that we only calculated the magnitude of the difference between A scores and B scores, without taking into account the direction of effect. Hence, all the randomization tests in this study were effectively two-sided.

Additionally, for each design type, data model, number of measurements, effect size, and test statistic, simulations were performed without missing data, to serve as a benchmark against which the missing-data methods could be evaluated. Overall, this meant $3 \times 3 \times 3 \times 3 \times 3 \times 2 = 486$ simulation conditions for each missing-data method and $3 \times 3 \times 3 \times 3 \times 2 = 162$ simulation conditions for the complete data, which added up to a total of $486 \times 3 + 162 = 1,620$ simulation conditions. For each condition, 100,000 datasets were simulated. In some of the simulation combinations, the number of possible randomizations was extremely high. For example, in an RBD with 40 measurements (20 blocks), the number of possible randomizations is $2^{20} = 1,048,576$. Hence, we decided to perform Monte Carlo RTs, where the reference distribution was constructed using 1,000 samples from all possible randomizations.

Customization of the missing-data methods

Due to the high percentage of missing data in some of the simulation conditions, it was possible that all data for one of the treatment levels could be missing for a particular simulated dataset. In such an exceptional case, it would be impossible to calculate the test statistic. The conservative nature of hypothesis testing dictates that the null hypothesis can be rejected only if the data provide sufficient evidence to reject the null hypothesis. The impossibility of calculating a test statistic from the available data can be considered a lack of evidence to reject. Hence, to make the analysis as conservative as possible, we decided to automatically conclude that we had failed to reject the

null hypothesis for that dataset. On the one hand, this strategy could be expected to control the probability of a type I error. On the other hand, this would possibly increase the probability of a type II error, hence reducing power. We tried to estimate the effects of this conservative strategy on type I errors and power, and found out that even in the most extreme case—for ABAB design data, 20 measurements, and 50% missing data—we were unable to calculate the test statistic in only 100 out of 100,000 cases. Hence, in the worst-case scenario, this would only result in a reduction of type I errors and power by approximately 0.1%.

For the randomized-marker method, it was also possible that all data points for a treatment level could be missing in any of the random assignments in the reference distribution. For such cases, we decided to automatically assume that the associated value of the test statistic was higher than the value for the observed data. Again, this ensured that we were consistent with our conservative approach, such that we always erred in favor of the null hypothesis, to ensure that the probability of a type I error did not increase.

While testing the TS-based SI method, we discovered that there was a small chance of the ARIMA fitting failing at higher missing-data percentages. Since a randomization test cannot be conducted without handling the missing data in some manner, a randomization test was conducted using a missing-data marker in those cases. The percentage of datasets in which the fitting failed was very low for our simulation conditions. For example, in a test involving the RBD, uniform data, 20 simulated measurements, an effect size of 2, and 50% missing data, the fitting failed only once in 100,000 repetitions. Hence, we were able to safely ignore the effect of such exceptions.

Programming and execution

The entire simulation study was programmed in R. One problem we encountered during early testing was that the simulation conditions involving NAP as the test statistic were taking

up to three times as much time as those involving MD as the test statistic. We were largely able to mitigate this problem by using C++ code to calculate NAP. In fact, using C++ code, the NAP calculation was even faster than the MD calculation.

Extrapolating from initial tests with 1,000 simulated datasets, we estimated a total computation time of approximately 20,000 h on a single-core processor running at 2.5 GHz. The simulations were run on Flemish Supercomputer Centre supercomputers. The nodes at the Flemish Supercomputer Centre utilized for this study consist of 20- or 24-core Intel processors running at 2.8-GHz or 2.5-GHz clock speed. Using a scheduler program, each of the 1,620 simulation conditions was scheduled to run on a separate core on these nodes.

Results

The results for type I errors (using effect size 0; see Table 2 and Fig. 1) showed that all three methods controlled type I errors effectively. Although the type I error rate breached the 5% level of significance in 128 out of 540 cases (23.70%; see Table 7 in the Appendix), the maximum was just 5.20%, so these can be attributed to simulation errors. The type I error rates for ABAB phase designs at 20 measurements and 50% missing data were particularly low. This can possibly be attributed to the conservative handling of cases in which all measurements corresponding to either A or B treatment levels were missing in a few of the possible randomizations. Since the number of measurements corresponding to a treatment level could be as low as six in our ABAB designs, with a small number of measurements and a large percentage of missing data, such scenarios were more likely to occur.

For an effect size of 1, the estimated power was the largest for the complete data, with power decreasing with increasing missing-data percentages (see Table 3 and Fig. 1). Furthermore, a consistent pattern was present in the results: The randomized-marker method performed closest to the complete data in all but a few cases, followed by MI. The SI

Table 2 Type I error rates, averaged over the entire set of data models, numbers of measurement choices, and test statistics

Design type	Complete data	Randomized marker			SI			MI		
		Missing-data percentage								
		10%	30%	50%	10%	30%	50%	10%	30%	50%
RBD	4.79	4.83	4.86	4.64	4.83	4.81	4.68	4.94	4.91	4.93
ABAB	4.88	4.80	4.52	3.97	4.84	4.79	4.75	4.85	4.89	4.87
MBD	4.98	4.98	4.99	4.96	4.99	5.00	5.01	4.98	4.97	5.01

ABAB, ABAB phase design; MBD, multiple-baseline design; MI, multiple imputation; RBD, randomized block design; SI, single imputation.

Fig. 1 Type I error rate and estimated power for different missing-data handling methods (and for complete data, for comparison), grouped by effect size (effect size 0 in row 1, effect size 1 in row 2, and effect size 2 in row 3), missing-data percentage (column 1), design type (column 2), and number of measurements (column 3), averaged over all other simulation conditions. *Complete* represents the complete data without missing

observations, *Marker* represents the randomized-marker method, *ES* represents effect size, *Missing* represents the percentage of missing data, and *N* represents the number of measurements. Please note the different y-axis ranges: 4–6 in row 1, and 0–100 in rows 2 and 3. ABAB, ABAB phase design; MBD, multiple-baseline design; MI, multiple imputation; RBD, randomized block design; SI, single imputation.

method rarely performed as well as the other two methods. The average power losses as compared to complete data, for which the average estimated power was 57.23%, were 16.46%, 23.59%, and 18.28% for the randomized-marker method, SI method, and MI method, respectively.

The estimated power values with an effect size of 2 followed the same pattern (see Table 3 and Fig. 1). We observed the highest power for complete data, with power decreasing with increasing missing-data percentages.

Again, the randomized-marker method performed best, closely followed by MI, and SI exhibited the worst results. The average power losses as compared to complete data, for which the average estimated power was 92.56%, were 9.87%, 22.39%, and 12.64% for the randomized-marker method, SI method, and MI method, respectively.

There were a few exceptions to these results. When the number of measurements was small, MI resulted in a

Table 3 Estimated power, averaged over the entire set of data models, numbers of measurement choices, and test statistics

Effect size	Design type	Complete data	Randomized marker			SI			MI		
			Missing-data percentage								
			10%	30%	50%	10%	30%	50%	10%	30%	50%
1	RBD	72.03	65.98	51.66	34.92	59.27	39.04	23.48	65.73	48.79	29.69
	ABAB	42.50	38.91	30.12	19.31	35.53	24.79	16.54	38.75	29.80	19.92
	MBD	57.17	52.91	42.38	30.80	48.81	33.68	21.71	52.64	40.30	24.95
2	RBD	99.21	98.50	94.46	81.31	91.75	71.72	47.89	98.27	91.95	72.27
	ABAB	85.01	82.01	71.93	52.51	78.92	61.64	40.06	81.62	71.12	52.28
	MBD	93.47	92.64	89.44	81.44	92.00	83.79	63.82	92.48	87.76	71.61

ABAB, ABAB phase design; MBD, multiple-baseline design; MI, multiple imputation; RBD, randomized block design; SI, single imputation.

larger estimated power than the randomized-marker method for AR1 data in an ABAB design (see Tables 8 and 9 in the Appendix). At larger percentages of missing data, there were also some cases (18 out of 108 cases for the ABAB design) in which SI performed better than the randomized-marker method. However, even for the ABAB design and AR1 data, when the number of measurements was higher (40, in our study), the randomized-marker method outperformed the other methods.

We also conducted an analysis of variance (ANOVA) of the power results. The results for effect size 0 were not included, since we were only interested in looking at estimated power. The results from the complete datasets were also not included, since we were interested in comparing the missing-data methods. The results from the ANOVA (see Table 4) indicated that only 2.5% of the explained variance could be attributed to the missing-data method, with effect size and missing-data percentage being the largest contributors. However, the results also indicated that the randomized-marker method, on average,

resulted in 2.3% more estimated power than did the MI method, and 9.8% more estimated power than the SI method (see Table 5).

Another aspect of RTs that is worth looking into, while comparing missing-data handling methods, is computation time. Comparing the simulation run times across the missing-data handling methods, we observed that on average the randomized-marker method was nearly twice as fast as SI and, depending on the type of design, five to ten times faster than MI (Table 6).

The results from the simulation study can be summarized as follows. First, all three methods controlled type I error rates close to or below the 5% level of significance for the RTs in the study. Second, due to the consistency of the results, we can reasonably conclude that the randomized-marker method compared favorably to SI and MI for the simulated conditions in the context of single-case RTs. Finally, the randomized-marker method was also significantly faster than the other two methods with respect to computation time.

Table 4 ANOVA of estimated power

Variable	Sum of squares	Percentage of explained variance
Design type	57,752	8.30
Data model	34,221	4.92
Test statistic	1,260	0.18
Effect size	385,062	55.35
No. of measurements	84,074	12.08
Missing-data method	17,110	2.46
Missing-data percentage	116,262	16.71
Residuals	121,028	

Table 5 Regression coefficients from ANOVA

Variable	Regression coefficient
Intercept	– 28.88
Design type: ABAB	0.00
Design type: MBD	14.30
Design type: RBD	17.83
Data model: AR1	0.00
Data model: Normal	14.53
Data model: Uniform	7.24
Test statistic: MD	0.00
Test statistic: NAP	– 2.28
Effect size	39.81
No. of measurements	1.14
Missing-data method: Randomized marker	0.00
Missing-data method: MI	– 2.30
Missing-data method: SI	– 9.82
Missing-data percentage	– 66.97

ABAB, ABAB phase design; AR1, autoregressive model with order 1 and standard normal errors; MBD, multiple-baseline design; MD, mean difference; MI, multiple imputation; NAP, nonoverlap of all pairs; RBD, randomized block design; SI, single imputation.

Discussion

In this study, we evaluated three different missing-data methods in terms of the type I error rate and estimated power of RTs. The results were quite convincing in favor of the randomized-marker method. These results provide a very strong basis for considering the randomized-marker method as the de facto missing-data handling method for single-case RTs.

However, as we mentioned previously, there were some exceptional cases in which the randomized-marker method performed poorly in comparison to SI and MI, particularly involving AR1 data in an ABAB design. We think

this was due to the use of preceding and succeeding measurements as predictors for imputation in the MI method and the use of an ARIMA model in the SI method, both of which predictably benefit from the large autocorrelation in AR1 datasets. We also think this exception was so prominent for ABAB phase designs probably because the simulated AR1 data for the ABAB phase design contained long series of data for a single phase, with no shift in values due to effect size applied; hence, the imputation methods were more successful. This probably indicates that when certain aspects of the observed data, including the presence of predictable patterns or of correlated covariate data, make imputation easier, imputation methods can outperform the randomized-marker method. However, when the number of measurements was 40, there were no exceptional cases. This indicates that when sufficient data are available, imputation in any form is not necessary, and the randomized-marker method is sufficient.

RTs are by nature computationally intensive (Edgington & Onghena, 2007). The computational complexity varies widely with the choice of design and test statistic. The time required for an RT increases with the number of measurements, as this increases the number of possible randomizations in an exact RT and increases the time required for calculating the randomizations and the test statistics. In our study, we used a Monte Carlo RT. Hence, we did not see any time variation due to increase in the number of possible randomizations. However, the time necessary to calculate a complex test statistic such as NAP multiple times can be a significant factor, too. We saw that the randomized-marker method holds a significant advantage over the other two methods in terms of computation time. Most of the time differences could be attributed to the differences in the missing-data handling methods, since all other steps remained the same. In this light, an additional argument can be made in favor of the randomized-marker method: that it does not add any

Table 6 Simulation run times (in seconds), averaged over the entire set of effect sizes, data models, numbers of measurement choices, and test statistics

Design type	Complete data	Randomized marker			SI			MI		
		Missing-data percentage								
		10%	30%	50%	10%	30%	50%	10%	30%	50%
RBD	11,416	6,036	5,714	5,352	10,842	10,696	11,085	50,653	50,567	50,595
ABAB	4,557	5,358	5,184	4,933	10,983	10,758	10,887	49,348	49,734	49,210
MBD	29,838	34,911	33,713	32,740	59,146	57,259	58,013	197,082	196,182	196,803

ABAB, ABAB phase design; MBD, multiple-baseline design; MI, multiple imputation; RBD, randomized block design; SI, single imputation.

additional time burden on an already computationally intensive inference method.

Limitations and future research

An obvious limitation to this simulation study is that the results are valid only for the simulation conditions we studied. It can be argued that the results could be approximated by interpolation to cover the entire range of effect sizes and missing-data percentages considered. However, the results cannot be generalized to data with other distributional characteristics. Particularly, in single-case research, the observed data in many experiments are discrete (e.g., counts or discrete ratings; Smith, 2012), as opposed to the continuous distributions used to simulate data in this study. We can expect power to be compromised for discrete distributions, due to the presence of ties in the data and in the reference distribution (Michiels et al., 2018). On the other hand, the RT procedure remains perfectly valid for discrete distributions, and this provides us an opportunity for future research into discrete data.

Another limitation to the generalizability of the results is that we used a Monte Carlo RT with 1,000 randomizations instead of an exact RT. This limitation was imposed mostly in order to manage computation time. As we mentioned previously, for an RBD with 40 measurements, the number of possible randomizations is 1,048,576. Trying to conduct an exact RT would have led to a thousand-fold increase in computation time for this design. Edgington (1969) argued that 1,000 randomizations in a Monte Carlo RT at a significance level of 5% is sufficiently precise. Michiels et al. (2018) calculated that if the expected power for a particular simulation condition is 80%, the 99% confidence interval for estimated power in a Monte Carlo RT with 1,000 randomizations is [77%, 83%].

A third major limitation in the scope of this study is the range of missing-data methods considered. A particular omission of interest is the EM algorithm, which is considered a popular method in the context of group designs (Smith et al., 2012). However, this is a parametric method, which relies on estimating model parameters and updating missing data values in each iteration. In the context of a fully nonparametric method such as RTs, this would not be appropriate. In fact, for the same reason, we decided not to explore the joint-modeling approach to MI. We also decided against exploring traditional simple SI methods, such as mean imputation, hot deck imputation, and so forth, in favor of a TS-based SI method, in spite of the former methods' prevalence, because their limitations are well known (Schafer & Graham, 2002).

Another limitation in this study is that we only used the “missing completely at random” mechanism to generate the missing data. This presents an opportunity for further research regarding the applicability of the randomized-marker method. Due to the assumptions made in an RT, the randomized-marker method may be suitable even when the missing data are missing at random (Little & Rubin, 2002) or not missing at random (Little & Rubin, 2002). In an RT, under the null hypothesis, it is assumed that the observed values (and missing-data markers) are exchangeable between treatment levels. Under this assumption, it becomes clear that the validity of the RT is not affected by the assignment of a missing-data marker to any treatment level, regardless of whether the missingness is dependent on the treatment level at the time of measurement or if missingness is dependent on the unrecorded actual value itself. However, the power of the RT might be affected substantially.

A final limitation of our study is that we only simulated univariate data. In SCEs, often multiple data streams are collected for the same unit. These data streams can be expected to be correlated, since they have the same source; hence, they can be utilized to impute missing values in the variable of interest. Multivariate methods such as vector autoregression can possibly be implemented in such scenarios. Covariate data can be particularly useful in MI, as has been demonstrated by Peng and Chen (2018). Exploring how correlated covariate data streams can be utilized to impute missing data will be an interesting approach for future research.

Conclusion

We can conclude that all three missing-data methods we tested result in adequate type I error rate control. However, the randomized-marker method outperformed SI and MI in terms of power in most of the simulation conditions. Additionally, the marker method is also faster. Hence, we recommend the randomized-marker method for handling missing data in RTs for SCEs, on the basis of our simulation results.

Acknowledgements This research was supported by an Asthenes long-term structural funding–Methusalem grant (number METH/15/011) by the Flemish Government, Belgium. The computational resources and services used in this work were provided by the Flemish Supercomputer Centre, funded by the Research Foundation–Flanders and the Flemish Government, department EWI.

Open Practices Statement The R code and results for this study are available on Open Science Framework at the following link: https://osf.io/bxw4s/?view_only=aebaa56338254c368148a7eca982154f. This study was not preregistered. However, the design of the simulation study was completely preplanned as part of the PhD project of T.K.D. at the KU Leuven.

Appendix

Table 7 Type I error rate at effect size 0

Design type	No. of measurements	Data model	Test statistic	Complete data	Randomized marker			SI			MI			
					Missing-data percentage									
					10%	30%	50%	10%	30%	50%	10%	30%	50%	
RBD	20	Uniform	MD	4.89	4.75	4.86	4.41	4.75	4.84	4.55	5.05	4.79	5.02	
			NAP	4.35	4.59	4.43	4.17	4.47	4.55	4.35	4.78	4.75	4.90	
		Normal	MD	4.84	4.85	4.83	4.43	4.90	4.94	4.52	4.81	4.88	4.91	
			NAP	4.42	4.58	4.62	4.10	4.64	4.45	4.20	4.93	4.78	4.79	
		AR1	MD	4.94	4.86	4.79	4.36	4.98	4.82	4.41	4.84	4.91	4.92	
			NAP	4.33	4.35	4.53	4.19	4.40	4.26	4.04	4.75	4.69	4.75	
	30	Uniform	MD	4.92	5.11	5.19	4.93	4.99	5.10	4.98	4.95	4.94	4.94	
			NAP	4.68	4.88	4.97	4.69	4.74	4.78	4.72	5.00	4.97	4.89	
		Normal	MD	5.03	4.99	4.97	4.90	4.98	4.96	4.85	4.95	5.02	5.00	
			NAP	4.78	4.75	4.89	4.55	4.78	4.81	4.56	4.82	4.93	4.95	
		AR1	MD	5.02	4.99	5.04	4.96	5.03	4.98	4.92	5.03	5.02	4.97	
			NAP	4.57	4.70	4.86	4.64	4.63	4.66	4.61	5.09	4.95	4.97	
	40	Uniform	MD	4.99	4.93	4.95	4.92	5.03	5.01	5.05	5.02	4.98	4.93	
			NAP	4.81	4.79	4.73	4.87	4.95	4.92	4.85	5.01	4.93	4.97	
		Normal	MD	5.04	4.95	4.99	4.93	5.05	4.97	5.01	4.98	5.03	5.00	
			NAP	4.87	4.98	4.92	4.82	4.99	4.88	4.85	4.90	5.04	4.98	
		AR1	MD	5.00	4.95	4.97	4.99	4.90	4.97	4.93	5.02	5.04	4.90	
			NAP	4.78	4.99	5.00	4.77	4.75	4.75	4.78	4.98	4.79	5.02	
	ABAB	20	Uniform	MD	4.59	4.56	4.08	2.99	4.66	4.67	4.44	4.74	4.62	4.72
				NAP	4.49	4.42	3.86	2.67	4.66	4.59	4.39	4.81	4.71	4.62
			Normal	MD	4.64	4.60	4.14	3.00	4.69	4.58	4.44	4.68	4.72	4.59
				NAP	4.58	4.45	3.92	2.72	4.57	4.45	4.40	4.61	4.75	4.63
			AR1	MD	4.73	4.63	3.87	2.82	4.61	4.62	4.51	4.69	4.77	4.70
				NAP	4.74	4.56	3.81	2.36	4.67	4.60	4.44	4.71	4.70	4.61
30		Uniform	MD	5.07	4.97	4.77	4.58	4.91	4.88	4.89	4.85	4.94	4.94	
			NAP	5.03	4.89	4.69	4.45	4.91	4.83	4.85	4.85	4.89	4.95	
		Normal	MD	4.95	4.99	4.75	4.43	4.98	4.89	4.86	4.96	5.06	4.92	
			NAP	4.96	4.92	4.71	4.32	4.91	4.87	4.83	4.95	5.05	4.96	
		AR1	MD	4.94	4.87	4.67	4.40	4.89	4.74	4.93	4.89	5.06	4.99	
			NAP	4.92	4.81	4.61	4.26	4.83	4.77	4.86	4.91	5.04	4.95	
40		Uniform	MD	5.12	4.94	4.97	4.71	5.08	4.97	4.99	4.89	4.93	5.00	
			NAP	5.09	4.89	4.88	4.78	4.86	4.82	4.94	5.05	4.90	4.98	
		Normal	MD	5.04	4.93	4.88	4.80	4.92	5.07	4.91	4.97	4.95	5.00	
			NAP	4.98	4.97	4.79	4.69	4.90	4.88	4.83	4.94	5.20	4.97	
		AR1	MD	5.00	5.00	4.91	4.70	5.02	5.06	5.06	4.92	4.91	5.06	
			NAP	4.90	4.99	5.03	4.76	5.06	4.93	4.98	4.89	4.88	5.03	
MBD		20	Uniform	MD	4.99	5.04	4.96	4.94	5.01	4.97	5.20	5.10	4.94	5.05
				NAP	5.02	4.92	4.99	4.87	4.92	5.02	5.13	4.95	4.93	4.97
			Normal	MD	4.93	5.04	5.00	4.85	4.93	4.93	5.08	5.05	5.01	5.06
				NAP	4.98	4.89	4.85	5.05	4.96	5.07	4.99	5.11	5.06	5.05
			AR1	MD	5.06	5.02	4.99	4.80	4.97	5.00	4.87	4.99	4.87	4.94
				NAP	4.92	4.92	5.05	4.83	5.13	5.01	4.92	4.89	4.98	4.99
	30	Uniform	MD	4.92	5.05	5.05	5.06	5.03	5.08	5.04	4.99	4.94	5.00	
			NAP	4.91	5.04	5.03	5.07	4.98	4.97	5.08	5.01	4.91	4.97	
		Normal	MD	5.07	4.75	4.97	4.80	4.84	5.01	4.94	4.87	4.95	5.00	
			NAP	5.05	4.72	4.95	4.92	4.81	4.91	4.95	4.91	4.87	4.97	
		AR1	MD	4.93	5.12	5.03	5.08	5.12	4.98	5.16	5.04	5.02	5.05	
			NAP	4.87	5.05	5.15	5.08	5.15	5.05	5.04	5.00	5.02	5.07	
	40	Uniform	MD	5.17	5.06	4.90	4.90	4.91	4.99	4.95	4.98	4.98	4.99	
			NAP	4.97	5.00	4.95	5.02	5.03	5.02	4.97	4.98	5.00	5.13	
		Normal	MD	5.00	5.03	5.03	4.90	4.97	4.97	4.94	4.94	5.00	5.11	
			NAP	4.95	5.06	4.98	5.07	5.02	4.92	4.98	5.00	5.08	4.89	
		AR1	MD	4.94	5.05	5.05	4.94	5.02	4.92	4.96	4.92	5.01	5.01	
			NAP	4.96	4.92	4.92	5.07	4.95	5.11	4.93	4.98	4.95	4.99	

Table 8 Estimated power at effect size 1

Design type	No. of measurements	Data model	Test statistic	Complete data	Randomized marker			SI			MI		
					Missing-data percentage								
					10%	30%	50%	10%	30%	50%	10%	30%	50%
RBD	20	Uniform	MD	48.28	43.11	31.57	18.65	29.38	15.36	8.28	42.20	29.11	16.56
			NAP	42.81	38.72	28.77	17.35	27.99	12.27	6.19	39.03	28.47	17.38
		Normal	MD	50.54	45.61	35.20	21.68	43.92	32.19	20.23	44.71	31.29	18.53
			NAP	46.72	42.39	32.30	20.01	40.59	28.10	17.39	42.80	31.59	19.89
		AR1	MD	70.19	59.00	39.11	21.21	57.71	35.76	19.44	58.11	34.22	16.59
			NAP	63.71	53.48	35.24	19.50	51.02	29.07	15.51	54.80	33.87	18.27
	30	Uniform	MD	71.65	65.50	52.28	35.92	43.79	22.98	13.61	64.40	48.05	29.12
			NAP	64.71	59.22	47.24	32.70	44.32	18.44	8.68	58.93	44.92	28.93
		Normal	MD	71.97	66.83	54.49	38.91	64.35	50.24	34.37	65.59	50.02	31.49
			NAP	68.79	63.91	51.45	36.00	60.64	44.04	28.37	63.59	49.44	32.83
		AR1	MD	89.48	80.55	59.00	37.49	80.48	56.07	33.08	80.84	54.71	27.58
			NAP	85.46	75.90	54.51	34.10	74.06	46.28	25.12	76.91	53.02	28.89
	40	Uniform	MD	85.02	80.59	68.34	50.13	56.60	30.96	17.72	79.32	63.92	42.37
			NAP	78.61	73.76	62.21	45.67	58.16	25.32	10.75	73.36	59.07	40.28
		Normal	MD	84.90	80.91	68.98	52.96	78.70	63.78	46.32	79.71	64.97	43.62
			NAP	82.76	78.66	66.50	49.49	75.49	57.50	38.25	77.85	63.94	44.65
		AR1	MD	96.47	91.15	73.30	50.19	91.78	72.10	45.00	91.71	69.91	38.48
			NAP	94.50	88.32	69.44	46.55	87.87	62.21	34.37	89.33	67.63	38.86
ABAB	20	Uniform	MD	33.89	29.48	19.74	9.67	23.72	13.55	8.41	29.86	20.61	12.13
			NAP	31.14	27.10	18.16	7.74	22.23	11.19	6.31	27.80	19.86	12.76
		Normal	MD	36.44	32.29	21.93	10.93	31.90	23.12	15.24	32.24	22.52	13.45
			NAP	34.71	30.61	20.40	8.67	30.27	21.59	13.80	31.20	22.69	14.47
		AR1	MD	21.31	19.54	14.71	8.01	19.57	15.73	11.41	19.73	15.77	11.32
			NAP	20.56	18.83	13.54	5.98	18.93	15.12	10.60	19.28	15.81	11.57
	30	Uniform	MD	54.97	50.01	38.03	23.65	38.89	22.15	13.19	49.31	35.93	21.75
			NAP	49.93	45.24	34.41	21.32	36.09	18.25	9.63	44.84	33.36	21.41
		Normal	MD	56.26	51.57	39.92	25.80	50.48	37.90	25.89	50.97	37.61	23.57
			NAP	54.00	49.35	37.75	24.08	47.74	34.33	22.57	49.33	36.91	24.36
		AR1	MD	25.40	24.24	20.83	15.64	23.87	20.35	16.25	23.70	20.77	15.99
			NAP	25.00	23.78	20.29	14.83	23.38	19.63	15.35	23.49	20.63	16.21
	40	Uniform	MD	68.14	63.14	49.79	33.62	50.47	29.89	16.93	62.13	47.43	29.82
			NAP	62.21	56.71	44.94	30.07	47.05	24.44	12.18	56.52	43.40	28.44
		Normal	MD	69.04	63.75	51.61	35.64	62.68	48.78	33.65	63.05	48.77	31.35
			NAP	66.94	61.65	48.92	33.75	59.77	44.02	29.06	61.38	48.06	32.02
		AR1	MD	27.65	26.72	23.78	19.30	26.70	23.34	19.19	26.34	23.20	19.16
			NAP	27.50	26.47	23.48	18.93	25.79	22.78	18.03	26.37	23.07	18.84
MBD	20	Uniform	MD	61.96	55.52	40.99	27.29	41.74	21.47	12.28	54.91	36.91	19.38
			NAP	53.49	47.13	34.27	22.33	37.42	17.40	8.68	47.21	32.87	18.79
		Normal	MD	62.80	56.02	42.13	28.74	55.31	40.26	25.95	55.73	37.94	20.59
			NAP	59.30	53.12	39.30	25.91	51.77	35.55	21.54	52.71	37.23	21.70
		AR1	MD	22.43	21.41	18.79	15.64	21.05	18.19	14.71	21.32	18.47	13.71
			NAP	21.21	20.32	17.54	14.18	20.23	16.95	12.97	20.34	17.75	13.33
	30	Uniform	MD	77.31	71.73	56.35	39.40	57.52	30.77	17.48	71.04	52.21	29.13
			NAP	69.48	63.65	48.46	32.87	53.34	25.31	13.04	63.11	46.06	26.82
		Normal	MD	77.81	71.48	56.71	40.23	70.94	54.58	36.13	70.74	52.71	30.41
			NAP	75.87	69.32	54.39	38.16	68.15	49.82	30.93	69.05	52.19	31.05
		AR1	MD	25.63	25.13	22.19	19.11	24.92	21.80	17.70	25.00	21.99	17.20
			NAP	24.88	24.24	21.32	18.03	24.18	20.77	16.32	24.36	21.43	16.86
	40	Uniform	MD	86.50	81.28	67.27	49.03	69.88	41.14	22.02	80.79	63.35	38.00
			NAP	80.44	74.40	59.10	41.91	65.17	34.45	15.87	74.00	56.65	34.17
		Normal	MD	86.49	81.44	67.56	50.06	81.37	66.16	45.34	80.91	63.85	38.32
			NAP	85.26	79.98	65.97	48.49	79.12	61.49	39.11	79.92	63.43	39.30
		AR1	MD	29.30	28.53	25.71	21.98	28.56	25.41	21.06	28.19	25.32	20.36
			NAP	28.88	27.61	24.84	21.03	27.91	24.68	19.73	28.16	24.95	20.03

Table 9 Estimated power at effect size 2

Design type	No. of measurements	Data model	Test statistic	Complete data	Randomized marker									SI			MI		
					Missing-data percentage														
					10%	30%	50%	10%	30%	50%	10%	30%	50%	10%	30%	50%	10%	30%	50%
RBD	20	Uniform	MD	98.27	96.51	87.67	59.81	64.63	32.84	14.85	95.32	79.41	45.70						
			NAP	95.17	92.50	82.00	54.53	72.01	27.57	9.64	92.27	78.80	50.30						
		Normal	MD	97.36	95.65	87.22	64.30	89.13	71.29	44.44	94.38	79.75	48.20						
			NAP	96.32	94.19	84.30	58.46	88.14	63.95	36.55	93.84	80.90	53.60						
		AR1	MD	99.85	98.59	87.88	59.33	96.10	75.91	44.22	97.86	79.26	42.17						
			NAP	99.50	97.39	84.61	55.15	94.08	66.23	35.20	97.21	80.63	47.85						
	30	Uniform	MD	99.97	99.90	99.08	92.25	77.91	49.62	28.42	99.84	97.29	78.55						
			NAP	99.67	99.27	96.93	86.93	92.87	45.03	15.97	99.28	96.01	79.98						
		Normal	MD	99.90	99.73	98.44	91.56	97.08	89.48	70.94	99.61	96.49	78.95						
			NAP	99.82	99.58	97.82	89.05	98.24	84.58	58.64	99.50	96.58	81.98						
		AR1	MD	100.00	99.96	98.42	87.29	99.71	93.56	70.73	99.96	96.46	70.64						
			NAP	99.99	99.90	97.52	84.02	99.59	87.20	56.61	99.92	96.29	74.24						
	40	Uniform	MD	100.00	100.00	99.94	98.72	84.78	60.56	38.35	99.99	99.73	93.64						
			NAP	99.98	99.95	99.50	96.40	98.55	59.79	22.05	99.94	99.36	93.32						
		Normal	MD	100.00	99.99	99.82	98.02	98.93	95.51	84.37	99.99	99.55	92.81						
			NAP	100.00	99.98	99.74	97.25	99.80	93.39	73.58	99.97	99.51	93.95						
		AR1	MD	100.00	100.00	99.81	96.00	99.98	98.48	85.23	100.00	99.57	86.69						
			NAP	100.00	100.00	99.63	94.55	99.98	95.89	72.20	100.00	99.45	88.26						
ABAB	20	Uniform	MD	83.12	76.12	55.10	26.50	63.53	34.71	16.66	75.60	55.83	30.85						
			NAP	79.52	72.40	50.36	19.85	61.14	28.86	10.83	73.19	54.82	32.54						
		Normal	MD	84.21	78.25	58.42	29.10	75.77	55.06	32.58	77.25	58.07	32.70						
			NAP	82.24	75.92	53.96	21.11	73.27	49.65	27.31	76.06	58.22	34.93						
		AR1	MD	61.02	56.69	42.59	22.24	55.38	41.87	26.34	55.77	43.38	26.93						
			NAP	59.23	54.31	37.97	14.39	53.43	39.03	23.04	54.53	42.80	27.81						
	30	Uniform	MD	98.27	96.73	89.55	67.70	90.18	62.62	35.28	96.13	85.79	60.40						
			NAP	96.53	94.46	85.95	62.87	88.27	55.66	26.39	94.19	83.70	60.62						
		Normal	MD	97.78	96.35	89.69	69.97	95.16	82.64	59.00	95.62	85.75	62.01						
			NAP	97.28	95.59	88.23	66.82	93.98	78.29	51.32	95.06	85.49	63.59						
		AR1	MD	71.01	68.45	61.28	46.93	67.26	57.81	42.36	67.44	58.37	43.90						
			NAP	69.90	67.17	60.09	43.92	66.01	55.57	38.63	66.62	58.11	44.50						
	40	Uniform	MD	99.78	99.46	97.27	86.09	97.41	80.39	49.47	99.31	95.09	78.09						
			NAP	99.26	98.62	95.11	81.90	96.46	74.41	39.49	98.49	93.47	77.05						
		Normal	MD	99.64	99.23	96.76	86.39	98.93	92.93	74.20	99.03	94.65	78.27						
			NAP	99.53	99.00	96.10	84.82	98.55	90.27	66.65	98.86	94.23	79.10						
		AR1	MD	76.41	74.09	68.56	57.84	73.44	65.67	52.31	73.14	66.31	53.93						
			NAP	75.47	73.37	67.69	56.72	72.48	64.11	49.32	72.85	66.14	53.82						
MBD	20	Uniform	MD	99.56	98.96	95.07	82.24	96.38	73.85	38.52	98.65	90.85	61.37						
			NAP	98.93	97.83	92.13	75.50	95.12	68.76	27.78	97.58	88.79	62.45						
		Normal	MD	99.41	98.80	94.70	82.29	98.39	90.74	67.83	98.47	90.73	62.36						
			NAP	99.32	98.56	94.03	80.13	97.87	87.45	58.30	98.21	90.76	65.31						
		AR1	MD	75.61	72.96	66.13	54.53	71.96	61.75	45.50	72.16	62.43	43.85						
			NAP	72.32	69.93	62.45	49.34	68.55	57.09	38.69	69.86	61.05	43.81						
	30	Uniform	MD	99.96	99.91	99.00	94.14	99.44	90.30	65.56	99.86	97.95	82.52						
			NAP	99.92	99.78	98.56	91.65	99.34	88.57	57.48	99.75	97.44	82.00						
		Normal	MD	99.95	99.86	98.78	93.87	99.82	97.81	85.61	99.81	97.81	82.62						
			NAP	99.96	99.87	98.93	94.04	99.81	96.98	80.85	99.82	98.11	84.37						
		AR1	MD	82.05	80.33	74.92	66.22	79.88	72.62	58.57	79.97	72.76	57.82						
			NAP	81.32	79.52	73.92	63.89	78.91	70.68	54.63	79.47	72.56	57.80						
	40	Uniform	MD	100.00	99.98	99.77	97.81	99.90	96.74	78.03	99.98	99.46	91.51						
			NAP	99.99	99.98	99.66	97.17	99.88	96.45	71.62	99.96	99.37	91.42						
		Normal	MD	100.00	99.97	99.68	97.50	99.97	99.44	93.31	99.97	99.39	91.48						
			NAP	100.00	99.99	99.76	98.19	99.98	99.28	90.81	99.97	99.54	92.89						
		AR1	MD	87.00	85.58	81.09	73.78	85.35	80.22	68.69	85.25	79.80	67.40						
			NAP	87.22	85.70	81.41	73.64	85.40	79.53	66.91	85.92	80.92	68.07						

References

- Adams, D. C., & Anthony, C. D. (1996). Using randomization techniques to analyse behavioural data. *Animal Behaviour*, 51, 733–738. <https://doi.org/10.1006/anbe.1996.0077>
- Anderson, B. D. O., & Moore, J. B. (1979). *Optimal filtering*. Englewood Cliffs, NJ: Prentice-Hall.
- Barlow, D. H., Nock, M. K., & Hersen, M. (2009). *Single case experimental designs: Strategies for studying behavior change* (3rd ed.). Boston, MA: Pearson/Allyn and Bacon.
- Box, G. E. P., & Jenkins, G. M. (1970). *Time series analysis: Forecasting and control*. San Francisco, CA: Holden-Day.
- Bulté, I., & Onghena, P. (2008). An R package for single-case randomization tests. *Behavior Research Methods*, 40, 467–478. <https://doi.org/10.3758/BRM.40.2.467>
- Busk, P. L., & Marascuilo, L. A. (1988). Autocorrelation in single-subject research: A counterargument to the myth of no autocorrelation. *Behavioral Assessment*, 10, 229–242.
- Chen, L.-T., Feng, Y., Wu, P.-J., & Peng, C.-Y. J. (2019). Dealing with missing data by EM in single-case studies. *Behavior Research Methods*. Advance online publication. <https://doi.org/10.3758/s13428-019-01210-8>
- Edgington, E. S. (1969). Approximate randomization tests. *Journal of Psychology*, 72, 143–149. <https://doi.org/10.1080/00223980.1969.10543491>
- Edgington, E. S., & Onghena, P. (2007). *Randomization tests* (4th ed.). Boca Raton, FL: Chapman & Hall/CRC.
- 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. <https://doi.org/10.3758/BRM.42.4.930>
- Ferron, J., & Onghena, P. (1996). The power of randomization tests for single-case phase designs. *Journal of Experimental Education*, 64, 231–239. <https://doi.org/10.1080/00220973.1996.9943805>
- Ferron, J., & Sentovich, C. (2002). Statistical power of randomization tests used with multiple-baseline designs. *Journal of Experimental Education*, 70, 165–178. <https://doi.org/10.1080/00220970209599504>
- Ferron, J., & Ware, W. (1995). Analyzing single-case data: The power of randomization tests. *Journal of Experimental Education*, 63, 167–178.
- Gardner, G., Harvey, A. C., & Phillips, G. D. A. (1980). Algorithm AS 154: An algorithm for exact maximum likelihood estimation of autoregressive-moving average models by means of Kalman filtering. *Journal of the Royal Statistical Society: Series C*, 29, 311–322. <https://doi.org/10.2307/2346910>
- Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science*, 8, 206–213. <https://doi.org/10.1007/s11121-007-0070-9>
- Guyatt, G. H., Jaeschke, R., & McGinn, T. (2002). Therapy and validity: N-of-1 randomized controlled trials. In G. Guyatt, D. Rennie, M. O. Meade, & D. J. Cook (Eds.), *Users' guides to the medical literature* (pp. 275–290). New York, NY: McGraw-Hill.
- Guyatt, G. H., Keller, J. L., Jaeschke, R., Rosenbloom, D., Adachi, J. D., & Newhouse, M. T. (1990). The n-of-1 randomized controlled trial: Clinical usefulness. Our three-year experience. *Annals of Internal Medicine*, 112, 293–299. <https://doi.org/10.7326/0003-4819-112-4-293>
- 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.
- Harvey, A. C., & Pierse, R. G. (1984). Estimating missing observations in economic time series. *Journal of the American Statistical Association*, 79, 125–131. <https://doi.org/10.1080/01621459.1984.10477074>
- Heyvaert, M., & Onghena, P. (2014). Randomization tests for single-case experiments: State of the art, state of the science, and state of the application. *Journal of Contextual Behavioral Science*, 3, 51–64. <https://doi.org/10.1016/j.jcbs.2013.10.002>
- 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. <https://doi.org/10.1177/0022466914525239>
- 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. <https://doi.org/10.1353/etc.2012.0011>
- Houle, T. T. (2009). Statistical analyses for single-case experimental designs. In D. H. Barlow, M. K. Nock, & M. Hersen (Eds.), *Single Case Experimental Designs: Strategies for Studying Behavior Change* (3rd ed., pp. 271–305). Boston, MA: Pearson/Allyn and Bacon.
- Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Fluids Engineering*, 82, 35–45. <https://doi.org/10.1115/1.3662552>
- Kazdin, A. E. (2011). *Single-case research designs: Methods for clinical and applied settings* (2nd ed.). New York, NY: Oxford University Press.
- Keller, B. (2012). Detecting treatment effects with small samples: The power of some tests under the randomization model. *Psychometrika*, 77, 324–338. <https://doi.org/10.1007/s11336-012-9249-5>
- Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). *Single-case design technical documentation*. https://ies.ed.gov/ncee/wwc/Docs/ReferenceResources/wwc_scd.pdf Accessed 12 Jul 2017
- Kratochwill, T. R., & Levin, J. R. (Eds.). (1992). *Single-case research design and analysis: New directions for psychology and education*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- Levin, J. R., Ferron, J. M., & Gafurov, B. S. (2018). Comparison of randomization-test procedures for single-case multiple-baseline designs. *Developmental Neurorehabilitation*, 21, 290–311. <https://doi.org/10.1080/17518423.2016.1197708>
- Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: John Wiley & Sons, Inc.
- Ma, H. H. (2006). An alternative method for quantitative synthesis of single-subject researches: Percentage of data points exceeding the median. *Behavior Modification*, 30, 598–617. <https://doi.org/10.1177/0145445504272974>
- Michiels, B., Heyvaert, M., & Onghena, P. (2018). The conditional power of randomization tests for single-case effect sizes in designs with randomized treatment order: A Monte Carlo simulation study. *Behavior Research Methods*, 50, 557–575. <https://doi.org/10.3758/s13428-017-0885-7>
- Moritz, S., & Bartz-Beielstein, T. (2017). imputeTS: Time series missing value imputation in R. *R Journal*, 9, 207–218.
- Onghena, P. (2005). Single-case designs. In B. Everitt & D. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (Vol. 4, pp. 1850–1854). <https://doi.org/10.1002/0470013192.bsa625>
- Onghena, P., & Edgington, E. S. (2005). Customization of pain treatments. *Clinical Journal of Pain*, 21, 56–68. <https://doi.org/10.1097/00002508-200501000-00007>
- Parker, R. I., & Vannest, K. (2009). An improved effect size for single-case research: Nonoverlap of all pairs. *Behavior Therapy*, 40, 357–367. <https://doi.org/10.1016/j.beth.2008.10.006>
- Peng, C.-Y. J., & Chen, L.-T. (2018). Handling missing data in single-case studies. *Journal of Modern Applied Statistical Methods*, 17, eP2488. <https://doi.org/10.22237/jmasm/1525133280>
- Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for

- improvement. *Review of Educational Research*, 74, 525–556. <https://doi.org/10.3102/00346543074004525>
- Roth, P. L. (1994). Missing data: A conceptual review for applied psychologists. *Personnel Psychology*, 47, 537–560. <https://doi.org/10.1111/j.1744-6570.1994.tb01736.x>
- Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York, NY: Wiley.
- Schafer, J. L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research*, 8, 3–15. <https://doi.org/10.1177/096228029900800102>
- Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, 7, 147–177. <https://doi.org/10.1037/1082-989X.7.2.147>
- 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. <https://doi.org/10.1177/074193258700800206>
- 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. <https://doi.org/10.3758/s13428-011-0111-y>
- Shumway, R. H., & Stoffer, D. S. (2017). *Time series analysis and its applications: With R examples* (4th ed.). New York, NY: Springer. <https://doi.org/10.1007/978-3-319-52452-8>
- Smith, J. D. (2012). Single-case experimental designs: A systematic review of published research and current standards. *Psychological Methods*, 17, 510–550. <https://doi.org/10.1037/a0029312>
- Smith, J. D., Borckardt, J. J., & Nash, M. R. (2012). Inferential precision in single-case time-series data streams: How well does the EM procedure perform when missing observations occur in autocorrelated data? *Behavior Therapy*, 43, 679–685. <https://doi.org/10.1016/j.beth.2011.10.001>
- 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. <https://doi.org/10.1177/0145445513510931>
- van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45, i03. <https://doi.org/10.18637/jss.v045.i03>
- 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. <https://doi.org/10.3758/BF03195492>
- Welch, W. J., & Gutierrez, L. G. (1988). Robust permutation tests for matched-pairs designs. *Journal of the American Statistical Association*, 83, 450–455. <https://doi.org/10.2307/2288861>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.