Behaviour Research and Therapy 119 (2019) 103414

Contents lists available at ScienceDirect

Behaviour Research and Therapy

journal homepage: www.elsevier.com/locate/brat

A multiple randomization testing procedure for level, trend, variability, overlap, immediacy, and consistency in single-case phase designs

René Tanious*, Tamal Kumar De, Patrick Onghena

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

ARTICLE INFO

Keywords:
Single-case experimental designs
ABAB phase design
Effect size measures
Randomization tests
False discovery rate
Visual analysis

ABSTRACT

We present an approach to draw multiple and powerful inferences for each data aspect of single-case ABAB phase designs: level, trend, variability, overlap, immediacy, and consistency of data patterns. We show step-by-step how effect size measures can be calculated for each data aspect and subsequently integrated as test statistics in multiple randomization tests. To control for Type I errors, we discuss three methods for adjusting the obtained *p*-values based on the false discovery rate: the multiple testing correction proposed by Benjamini and Hochberg (1995), the adaptive correction suggested by Benjamini and Hochberg (2000), and the correction taking into account the dependency between the tests (Benjamini & Yekutieli, 2001). We apply this approach to a published data set and compare the results to the conclusions drawn by the authors based on visual analysis. The multiple randomization testing procedure can give more detailed information about which data aspects are affected by the single-case intervention. We provide generic R-code to execute the presented analyses.

1. Introduction

Intervention effects in single-case experimental designs (SCEDs) are still predominantly established through visual analysis although there is an upward trend in the inclusion of statistical analyses (Kratochwill & Levin, 2014; Shadish, Hedges, & Pustejovsky, 2014; Smith, 2012). Accepted guidelines on the collection and analysis of single-case data recommend visually inspecting six data aspects: level, trend, variability, immediacy, overlap, and consistency (e.g., Kratochwill et al., 2010; Kratochwill et al., 2013). Following Horner et al. (2005), level is defined as the mean score on the dependent variable in a given phase; trend refers to the slope of the best fitting straight line in a phase; variability refers to the degree of fluctuation of the scores on the dependent variable; immediacy assesses the duration between the introduction or withdrawal of the intervention and any observed effects; and overlap refers to the proportion of data points on the dependent variable overlapping between adjacent phases. Consistency refers to the degree of consistency between data patterns from phases implementing the same condition (Kratochwill et al., 2010).

A systematic review of 119 applied single-case ABAB studies indicated that 73% of studies relied solely on visual analysis (Tanious, De, Michiels, Van den Noortgate, & Onghena, 2018). While it has been argued that visual analysis is sensitive to abrupt as well as subtle changes over time (Lane & Gast, 2014), numerous studies have also pointed out considerable drawbacks of visual analysis. Briefly, these drawbacks include low interrater agreement (Heyvaert & Onghena, 2014); a lack of clear decision rules (Perdices & Tate, 2009); serial dependency in the data misguiding visual judgement (Matyas & Greenwood, 1990); and high Type II error rates (Ottenbacher, 1990). In light of these shortcomings, recent guidelines recommend employing visual and statistical analyses in a complementary way (Tate et al., 2016a). Given the predominant role of visual analysis in the field, multiple efforts have been made to operationalize visual analysis through systematic protocols. For example, Barton, Lloyd, Spriggs, and Gast (2018) proposed a seven step visual analysis protocol for assessing the presence of a functional relation and Maggin, Briesch, and Chafouleas (2013) developed a visual analysis protocol specifically following the What Works Clearinghouse guidelines. These efforts are an important step forward in the methodology and data analysis of SCEDs.

The present paper demonstrates an analytical approach designed to supplement visual analysis of each data aspect with quantifications and information about statistical significance. As Parker and Vannest (2009) point out, an approximate visual judgement of intervention effects in SCEDs may suffice for low-stakes, in-house decisions, but “measuring the amount of change with known precision is best accomplished by an effect size index” (p. 358). Furthermore, effect size measures complement visual analysis with an objective quantification (Manolov, 2018).

* Corresponding author. Faculty of Psychology and Educational Sciences, KU Leuven Tiensestraat 102 box 3762, B-3000 Leuven, Belgium.
E-mail addresses: rene.tanious@kuleuven.be (R. Tanious), tamalkumar.de@kuleuven.be (T.K. De), patrick.onghena@kuleuven.be (P. Onghena).

https://doi.org/10.1016/j.brat.2019.103414
Received 13 February 2019; Received in revised form 16 May 2019; Accepted 28 May 2019
Available online 05 June 2019

Acknowledging the complementarity of visual and statistical analyses, many effect size measure for SCEDs have been developed over the past decades. In 1987, Scruggs, Mastropieri, and Casto proposed the percentage of nonoverlapping data (PND) as the first effect size measure to quantify intervention effects in SCEDs. While overlap has certainly received the most attention, 30 years after Scruggs et al.'s proposal all six data aspects have been operationalized (cf. Tanious, De, Michiels, Van den Noortgate, & Onghena, 2019). It should be noted that the data aspects level, trend, and variability were originally proposed as *within*-phase measures: "In addition to comparing the level, trend, and variability of data within each phase, the researcher also examines data patterns across phases by considering the immediacy of the effect, overlap, and consistency of data in similar phases" (Kratochwill et al., 2010, p. 18). To measure the effect of the intervention on the dependent variable, effect size measures for these data aspects can be adapted to quantify changes between adjacent baseline and experimental phases in level, trend, and variability. Overlap and immediacy are per definition *between*-phase measures because they can only be quantified using data from two adjacent phases. Consistency also falls in the between-phase category because it is assessed using data from two phases, but differs from overlap and immediacy in one respect. Rather than using data from adjacent phases, consistency is assessed between experimentally similar phases, e.g., both baselines phases in an ABAB design. A major advantage of these quantifications for each data aspect is that they can be compared across studies. For example, in group studies Cohen's *d* is a popular effect size metric to quantify an intervention effect between a treatment and a control group which can be compared across studies and included in meta-analyses. Following this logic, Hedges, Pustejovsky, and Shadish (2012, 2013) and Pustejovsky, Hedges, and Shadish (2014) developed a *d* metric for SCEDs for level as the data aspect. An overview of the aims, methodology and underlying assumptions of effect sizes for SCEDs as well as the correct reporting of them is available in Shadish, Hedges, Horner, and Odom (2015). Shadish et al. (2015) distinguish between *between*-case and *within*-case effect size measures for SCEDs. While *between*-case effect size measures are rather rare, they can have great added value in meta-analyses, for example, the *d* statistic proposed by Hedges et al. (2012, 2013) and Pustejovsky et al. (2014) for finding an average treatment effect across participants. The majority of effect size measures for SCEDs are *within*-case measures that contrast measurements taken under one condition with measurements taken under a different condition for a single entity. These effect size measure can then be descriptively compared across participants.

The analytical approach presented in this paper utilizes one *within*-case quantification per data aspect to express the intervention effect on that feature of the data. These quantifications are then incorporated as test statistics in multiple randomization tests to derive *p*-values. Finally, the *p*-values are corrected for multiple testing. Before turning to an in-depth discussion of the presented methodology, three considerations underlying the approach should be addressed: 1) Why should visual analysis and effect size calculation be supplemented with significance tests? 2) Which methods exist to perform significance tests in an SCED framework? 3) If multiple tests are performed, how can we account for possible false positives?

Regarding the first question, it is important to note that each technique gives distinctive information. Visual analysis can be used to monitor the data session-to-session during a single-case study and modify the intervention if necessary (Lane & Gast, 2014). Furthermore, visual analysis can give a conservative estimate of intervention effectiveness (Park, Marascuilo, & Gaylord-Ross, 1990). Effect size calculation supplements this estimate by expressing the exact size of the effect (Parker & Vannest, 2009) and enabling comparison of effects between studies (Wilkinson, 1999). Significance tests can be employed to ascertain if any observed effect differs from a null effect. The obtained *p*-value expresses the probability of obtaining the observed data given the treatment has no effect (e.g., Castro Sotos, Vanhoof, Van den Noortgate, & Onghena, 2007; Moore & McCabe, 2005).

Regarding the second question, one can broadly distinguish between parametric and non-parametric tests. It has repeatedly been argued that SCED datasets often fail to meet parametric assumptions (e.g., Ferron, Foster-Johnson, & Kromrey, 2003; Parker & Vannest, 2009; Weaver & Lloyd, 2018). Taking repeated measurements of a single case over a period of time results in various degrees of autocorrelation in SCEDs (Busk & Marascuilo, 1988; Olive & Smith, 2005; Smith, 2012; Solomon, 2014). This data structure results in violations of critical assumptions of parametric tests, i.e., data sampled from a normally distributed population and independent error structure. The use of parametric statistical tests such as *F*- or *t*-tests for SCEDs is therefore generally not recommended (Brossart, Parker, Olson, & Mahadevan, 2006; Bulté & Onghena, 2013; Center, Skiba, & Casey, 1985-86; Heyvaert, Wendt, Van den Noortgate, & Onghena, 2015; Levin, Ferron, & Kratochwill, 2012; Rvachew, 1988). A particular class of non-parametric tests well suited for SCEDs are randomization tests as they do not rely on any assumption regarding the sampling distribution. Instead, randomization tests derive their own sampling distribution and obtain their validity from the randomization procedure actually used in the experiment. A second major advantage of randomization tests is their versatility (Onghena, 1992). Randomization tests can incorporate various kinds of test statistics according to the researcher's interest, e.g., mean differences or difference in trend lines between adjacent phases. We will turn to a detailed discussion of randomization tests in the methodology section.

Regarding the third question, if multiple tests are conducted in the same study the Type I error rate might deviate from the nominal level. However, several techniques exist to reduce the risk of finding false positives. The multiple testing corrections discussed in the present study are based on the false discovery rate (FDR; Benjamini & Hochberg, 1995). As the six data aspects in an SCED study are defined beforehand, the presented analysis prevents conducting significance tests only on selected data aspects for which the researcher expects favorable outcomes. For an in-depth discussion of these issues the interested reader is referred to Simmons, Nelson, and Simonsohn (2011) and Wicherts et al. (2016). In the following sections, we present each quantification used in the present study and then turn to an explanation of randomization tests.

2. Methods for quantifying each data aspect

Table 1 gives an overview of the quantifications used in this study with key references for further reading. As the presented technique is

Table 1
Selected quantifications for each data aspect with key reference(s).

Data aspect	Effect size measure	Key reference(s)
Level	Pooled standardized mean difference (pSMD)	Cohen (1992); Beretvas and Chung (2008)
Trend	Ordinary least squares comparison (OLS)	Kromrey and Foster-Johnson (1996)
Variability	Variance ratios (VR)	Kromrey and Foster-Johnson (1996)
Overlap	Nonoverlap of all pairs (NAP)	Parker and Vannest (2009)
Immediacy	Immediate treatment effect index (ITEI)	Michiels, Heyvaert, Meulders, & Onghena (2017)
Consistency	Consistency of data patterns (CONDAP)	Tanious, De, Michiels, Van den Noortgate, & Onghena (2019a)

generic, researchers may also use other quantifications as test statistics to perform the randomization tests.

2.1. Level: pooled standardized mean difference

The pooled standardized mean difference (pSMD) for SCEDs is an adaptation of the widely used group-based Cohen's *d* (Cohen, 1992). In the framework of Cohen, the mean of the control group is subtracted from the mean of the experimental group and divided by the pooled standard deviation of the two groups. The pooled variance is derived by adding up the variances –adjusted for the degrees of freedom- divided by *N*-2. Taking the square root gives the pooled standard deviation. Following this logic, the pSMD can be defined as shown in Equation (1):

$$pSMD = \frac{M_B - M_A}{\sqrt{\frac{(n_B - 1) * SD_B^2 + (n_A - 1) * SD_A^2}{n_B + n_A - 2}}} \quad (1)$$

In the context of SCED phase designs, the mean of the baseline phase (M_A) is subtracted from the mean of the experimental phase (M_B). The pooled standard deviation is derived by adjusting each standard deviation for the number of measurement occasions in that phase, divided by the total number of measurement occasions. Thus, contrary to Cohen's *d,* the number of participants is replaced by the number of measurement occasions.

2.2. Trend: ordinary least squares comparison

The ordinary least squares (OLS) comparison approach suggested by Kromrey and Foster-Johnson (1996) separately fits regression lines to A and B phases and obtains an effect size measure from the associated R^2 values, a statistic familiar to applied researchers. In the first step, a regression line is calculated as if no change in slope occurred (Equation (2)). Subsequently, a regression line is calculated in which the slopes for each regression line are different (Equation (3)). The obtained effect size measure called f^2 can be calculated as shown in Equation (4) where R_2^2 and R_3^2 are the coefficients of determination of Equations (2) and (3).

$$y = b_0' + b_1'T + b_2'X + e \quad (2)$$

$$y = b_0'' + b_1''T + b_2''X + b_3''XT + e \quad (3)$$

$$f^2 = \frac{R_3^2 - R_2^2}{1 - R_3^2} \quad (4)$$

In the regression equations (2) and (3) T represents the time regressor defined as the consecutive moments of observation with the first measurement occasion taking the value 0. X is a dummy coded variable taking the value 0 for the A-phase measurements and the value 1 for the B-phase measurements. The additional regression weight b_3 in Equation (3) allows to fit separate regression lines due to the interaction between the dummy coded variable and the time regressor.

2.3. Variability: variance ratios

The variance ratio (VR) gives information about the change in variability of the target behavior after introducing the intervention. The first step to calculating the VR is to calculate the variance for each experimental phase separately. The variance ratio is obtained by dividing the larger variance by the smaller variance (Kromrey & Foster-Johnson, 1996) as shown in Equation (5).

$$VR = \frac{S_L^2}{S_S^2} \quad (5)$$

2.4. Overlap: nonoverlap of all pairs

The nonoverlap of all pairs (NAP) "summarizes data overlap between each phase A datapoint and each phase B datapoint, in turn. A nonoverlapping pair will have a phase B datapoint larger than its paired baseline phase A datapoint. NAP equals the number of comparison pairs showing no overlap, divided by the total number of comparisons" (Parker & Vannest, 2009, p. 358). This definition can be conceptualized as shown in Equation (6).

$$\frac{(N_A \times N_B) - \sum \left[\left(Overlap_{A_{ij}B_{ij}} \times 1 \right) + \left(Tie_{A_{ij}B_{ij}} \times 0.5 \right) \right]}{N_A \times N_B} \quad (6)$$

An overlapping pair receives a score of one and a tie (two scores having the same value) receives a score of 0.5. This sum is subtracted from the total number of comparisons $N_A \times N_B$ and then divided by the total number of comparisons to obtain the NAP.

2.5. Immediacy: immediate treatment effect index

The immediate treatment effect index (ITEI; Michiels, Heyvaert, Meulders, & Onghena, 2017) assesses whether any change in the dependent variable occurs abruptly after introducing the intervention. Following the definition of Kratochwill et al. (2010), Michiels et al. (2017) conceptualized the ITEI as shown in Equation (7).

$$ITEI = \frac{\sum_{i=n-3}^{n_A} A_i}{3} - \frac{\sum_{i=nA+1}^{nA+3} B_i}{3} \quad (7)$$

This equals the mean of the last three phase A data points minus the mean of the first three phase B data points.

2.6. Consistency: consistency across data patterns of similar phases

The consistency across data patterns (CONDAP) measure is a Manhattan distance based measure that assesses the similarity of two data patterns (Tanious et al., 2019). Contrary to the other five data aspects, CONDAP is not calculated to assess changes between different phases. The primary goal of CONDAP is to assess whether the introduction and withdrawal of the intervention leads to similar patterns of responding for similar phases. In the context of an ABAB design, this means that CONDAP assesses the similarity of the A1 and A2 data patterns on the one hand and the B1 and B2 data patterns on the other hand. CONDAP can be calculated as shown in equation (8) where k equals the number of paired sequences; n_s equals the number of measurement occasions in the shorter phase; x_{ij} equals the i_{th} measurement in the j_{th} sequence of A1 or B1; y_{ij} equals the i_{th} measurement in the j_{th} sequence of A2 or B2;

n_x equals the number of data points in A1 or B1; and n_y equals the number of data points in A2 or B2.

$$CONDAP = \frac{\frac{1}{kn_s} \sum_{j=1}^{k} \sum_{i=1}^{n} \left| x_{ij} - y_{ij} \right|}{\sqrt{\frac{(n_x - 1) * SD_x^2 + (n_y - 1) * SD_y^2}{n_x + n_y - 2}}} \quad (8)$$

The numerator calculates the average Manhattan distance between the two data patterns. Each A1 data point is compared to the corresponding A2 data point. The Manhattan distance is the distance between these two data points on the y-axis. This calculation is carried out for each pair of data points between the two phases. If the phases differ in lengths, the shorter phase is compared to each possible sequence of equal length in the longer phase. The obtained mean Manhattan distance is divided by the pooled standard deviation of the two phases being compared. Lower CONDAP values indicate higher consistency and vice versa.

Generic R-code to carry out all calculations presented in this article

–including the above quantifications- is freely available in the online supplementary material and upon request to the first author.

3. Randomization tests for SCEDs

Randomization tests are a form of non-parametric significance tests that have added great value to the methodological toolbox of SCED analytical techniques as it has repeatedly been argued that SCED datasets often fail to meet parametric assumptions (e.g., Ferron et al., 2003; Parker & Vannest, 2009; Weaver & Lloyd, 2018). Contrary to other popular non-parametric tests used in group studies (e.g, Kruskal-Wallis and Wilcoxon rank tests), randomization tests do not require the conversion of scores to ranks. Randomization tests operate on the principle that "in experiments in which randomization is performed, the actual arrangement of treatments […] is one chosen at random from a predetermined set of possible arrangements" (Welch, 1937, p. 47). In general, the *p*-value obtained by performing a randomization test for SCEDs is the likelihood of obtaining the observed data given that there is no differential effect of the levels of the independent variable (Onghena & Edgington, 2005). Recognizing the potential of randomization tests for analyzing data from SCEDs, Edgington (1967) first proposed randomization tests for alternating treatment designs and later extended his work to applications for AB and ABA designs (Edgington, 1975a).

Following Edgington's line of thought, Onghena (1992) proposed a randomization procedure for the ABAB design. As Ferron et al. (2003) summarized:

"Onghena (1992) illustrated an alternative way to incorporate randomization into reversal designs. A researcher who wishes to conduct an ABAB design can randomly choose a triplet of intervention points (i.e, change from Al to Bl [t1], change from Bl to A2 [t2], change from A2 to B2 [t3]) such that all phases are represented and have at least a certain number of observations." (p. 269).

A comprehensive in-depth discussion of randomization tests in general, their rationale, and applications to several SCEDs is offered by Edgington and Onghena (2007). The following discussion of the generic steps involved in conducting a randomized SCED and performing the subsequent randomization test is based on Heyvaert and Onghena (2014). A step-by-step numerical example executing the generic steps to an applied data set will be given in the methodology section.

In a first step, the null hypothesis and alternative hypothesis are defined, the level of significance α is chosen, and the researcher determines the total number of measurement occasions. Before conducting the experiment, the researcher should furthermore select the test statistic(s) that conform to the expected effect of the treatment. In the presented approach the test statistics are the six quantifications presented in the previous section.

In a second step, the researcher determines the randomization scheme. The randomization scheme might either be restricted or unrestricted (cf. Edgington, 1980). In the unrestricted case, the three phase change moments might occur after any measurement occasion randomly. With an unrestricted randomization scheme, the researcher might however end up with a research design that does not meet quality standards (e.g., less than three data points for one or several phases) (Kratochwill et al., 2010). With a restricted randomization scheme, the introduction or withdrawal of the treatment can be randomly determined after at least three measurement occasions within each phase. Additional factors might be included as well in the restricted randomization scheme (e.g., the stability of the baseline or the variability of the emerging data pattern) (cf. Heyvaert et al., 2015).

In a third step, the researcher randomly chooses one of the possible assignments. Using this assignment, the experiment is then carried out and the test statistics are calculated.

In a fourth step, the test statistics are calculated for each of the possible assignments. The obtained distribution of test statistics forms the reference distribution (Manolov & Onghena, 2018; Michiels & Onghena, 2018) and its function is identical to the function of the sampling distribution in between-groups studies (Heyvaert & Onghena, 2014). Heyvaert and Onghena (2014) furthermore make an important point for obtaining the reference distribution:

> "In order to construct the randomisation distribution only the random assignment as it was actually implemented while carrying out the study is taken into account, because this assignment procedure is the stochastic foundation for the reference distribution of an RT [randomization test]. In all possible assignments for which the test statistic is computed, the exact sequence of the measurements actually obtained remains intact; only the levels of the independent variable (the "treatment labels") are redistributed or reshuffled to generate the virtual replications." (p. 509)

The *p*-value of a randomization test equals the proportion of test statistics in the reference distribution that are as extreme or even more extreme as the obtained test statistic from the actual experiment.

4. Method

In this paper, we illustrate with an applied data set how the previously discussed quantifications can be integrated as test statistics in multiple randomization tests. We present the obtained test statistic and *p*-value for each data aspect. Next, we outline how the obtained *p*-values can be corrected for multiple testing to control for false positives (Type I error rate) using three different multiple testing corrections. Finally, we compare the results of our analyses with the conclusions reached by the original authors based on a visual analysis of the graphed data.

4.1. An example of an ABAB design

The dataset we use to illustrate the multiple randomization test procedure was obtained from Feeney and Ylvisanker (2003). The authors used an ABAB phase design to investigate the effects of a multicomponent cognitive-behavioral intervention on the challenging behavior of two young children with traumatic brain injury. Fig. 1 shows the results of Mark, a 7-year old boy, on the first dependent variable: the number of occurrences of challenging behavior in the classroom. This variable was measured on 28 days under two conditions. During A-phases, Mark's behavior was observed under pre-existing classroom conditions. During B-phases, several supports were added to Marks's daily routine. The raw data were retrieved from the published graph using "GetData Graph Digitizer" version 2.26 (Fedorov, 2013).

As can be seen in Fig. 1, all phase B scores are smaller than the data points in each preceding A-phase. Regarding the dependent variable displayed in Fig. 1, the authors conclude that:

Hypotheses 1. [*sic*] (that the challenging behaviors would decrease in frequency) […] was confirmed for both children. The reversal (ABAB) design clearly demonstrated that the intervention had the effect of dramatically reducing the frequency […] of the challenging behaviors to levels considered acceptable in the students' integrated community school placements. (Feeney & Ylvisanker, 2003, pp. 45–46)

In the remainder of the present article we will use the data from Fig. 1 to apply the multiple randomization test procedure described in the previous section. To facilitate interpretation of the design, analyses, and results, we followed the authors' description of the study as a "single-subject reversal (ABAB) design […] to evaluate the effectiveness of the multicomponent intervention" (Feeney & Ylvisanker, 2003, p. 40). For the sake of completeness, it should be noted that the first B condition included the criterion that the frequency of negative behavior may not occur more than twice per day for five consecutive days before returning to baseline measures. It should furthermore be noted that the study was published in 2003 and that guidelines for conducting and analyzing SCEDs have changed since. For example, both baseline

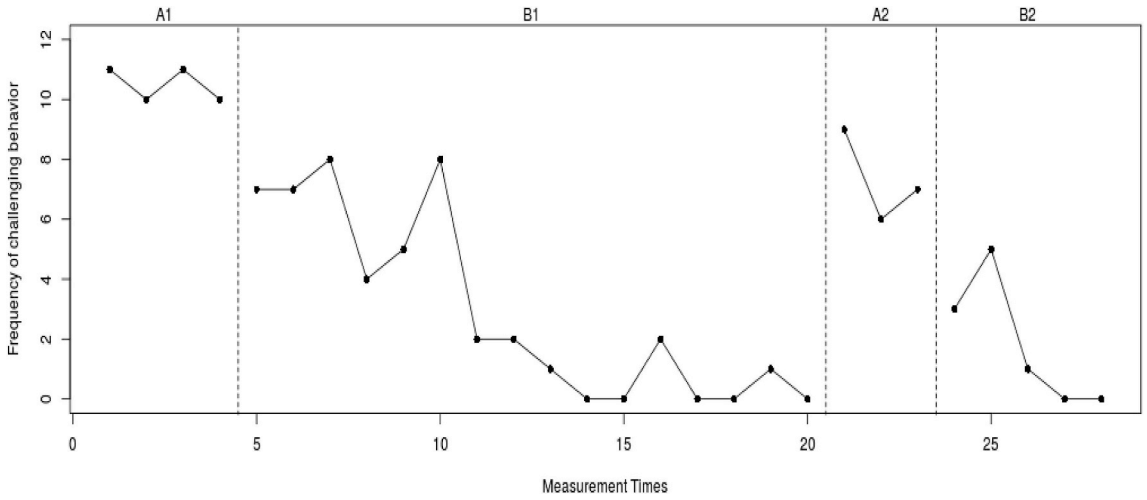

Fig. 1. Results from an ABAB design to decrease the frequency of challenging behavior in a child with traumatic brain injury. Data from Feeney, T. J., & Ylvisanker, M. (2003). Context-sensitive behavioral supports for young children with TBI: Short-term effects and long-term outcome. *The Journal of Head Trauma Rehabilitation, 18*, 33–51.

phases in the example data set contain less than five data points. Recent standards, however, require a minimum of five data points per phase to meet evidence standards without reservation (Ganz & Ayres, 2018; Horner et al., 2005; Kratochwill et al., 2013, 2010; Tate et al., 2016b; U.S. Department of Education, 2016), while less than five data points per phase can only meet minimum evidence standards (e.g., Beeson & Robey, 2006; Kratochwill et al., 2010).

Step 1: Hypothesis formulation, significance level, and number of measurements

The first step has to be executed independently of the data to warrant statistical conclusion validity. This can for example be done *a priori* before conducting the experiment. If (parts of) the first step are not executed *a priori*, alternative ways of data analyses have been proposed using masked graphs (e.g., Ferron & Foster-Johnson, 1998; Ferron & Jones, 2006). The alternative hypothesis should be formulated in terms of the expected effect. A first basic consideration is to evaluate whether we expect the data in the experimental phase to differ in a specific direction from the data in the preceding baseline phase. In the example data set, the researchers expected the intervention to lead to a decrease in the frequency of challenging behaviors. Therefore, we choose this as our one-sided alternative hypothesis. If the direction of the effect cannot be estimated beforehand, a two-sided alternative should be chosen. The general null hypotheses is that the treatment is ineffective, i.e. that it does matter whether any given measurement was taken during an A- or B-phase. As our significance level, we choose the conventional $\alpha = .05$. In the example data set, 28 measurements were taken. For the validity of the randomization procedure, it is an important assumption that this number had been determined before the start of the experiment.

Step 2: Determination of the randomization scheme

Following quality guidelines for conducting SCEDs (Kratochwill et al., 2010), we choose a restricted randomization scheme which allows for at least three measurements per phase. There are $\binom{28 - 3(3+1) + 3}{3} = 969$ possible assignments that allow for at least three measurements per phase given the total number of 28 measurements (not all assignments are listed):

AAABBBAAABBBBBBBBBBBBBBBBB
AAAABBBAAABBBBBBBBBBBBBBBB
AAAAABBBAAABBBBBBBBBBBBBBB
AAAAAABBBAAABBBBBBBBBBBBBB
…
AAAAAAAAAAAAABBBBBBAAAABBBBB
AAAAAAAAAAAAABBBBBBBAAAABBBB
AAAAAAAAAAAAABBBBBBBAAAAABBB
AAAAAAAAAAAAABBBBBBBAAAAAABB

Step 3: Conducting the experiment and calculating the observed test statistic

To conduct the experiment, the researcher chooses one of the 969 possible assignments at random. The assignment used in the example data set was AAAABBBBBBBBBBBBBBBBAAABBBBB and the obtained values for each measurement were 11,10,11,10,7,7,8,4,5,8,2,2,1,0,0,2,0,0,1,0,9,6,7,3,5,1,0,0 as visualized in Fig. 1. Using the raw data, we calculated the observed test statistic for each data aspect using the quantifications presented in Table 1. Equations (9)–(11) show as an example the calculation of the observed test statistic for immediacy.

$$ITEI_{A1B1} = \frac{10 + 11 + 10}{3} - \frac{7 + 7 + 8}{3} = 3.00 \quad (9)$$

$$ITEI_{A2B2} = \frac{9 + 6 + 7}{3} - \frac{3 + 5 + 1}{3} = 4.33 \quad (10)$$

$$Ave_{ITEI} = \frac{3.00 + 4.33}{2} = 3.67 \quad (11)$$

Three generic steps are involved in the calculation of the observed test statistic for level, trend, variability, overlap, and immediacy. Firstly, a quantification is calculated for changes from A1 to B1 (Equation (9)). Subsequently, a quantification is calculated for changes from A2 to B2 (Equation (10)). Finally, the average of the two obtained quantification is taken (Equation (11)). The ITEI for the first change from baseline to intervention equals 3.00 and for the second change from baseline to intervention 4.33. On average, the introduction of the intervention thus leads to an immediate decrease of 3.67 occurrences of challenging behavior. As consistency is not compared between adjacent phases, the first two generic steps have to be modified slightly. Firstly, CONDAP is calculated for the two baseline phases. Subsequently, CONDAP is calculated for the two intervention phases. The third step

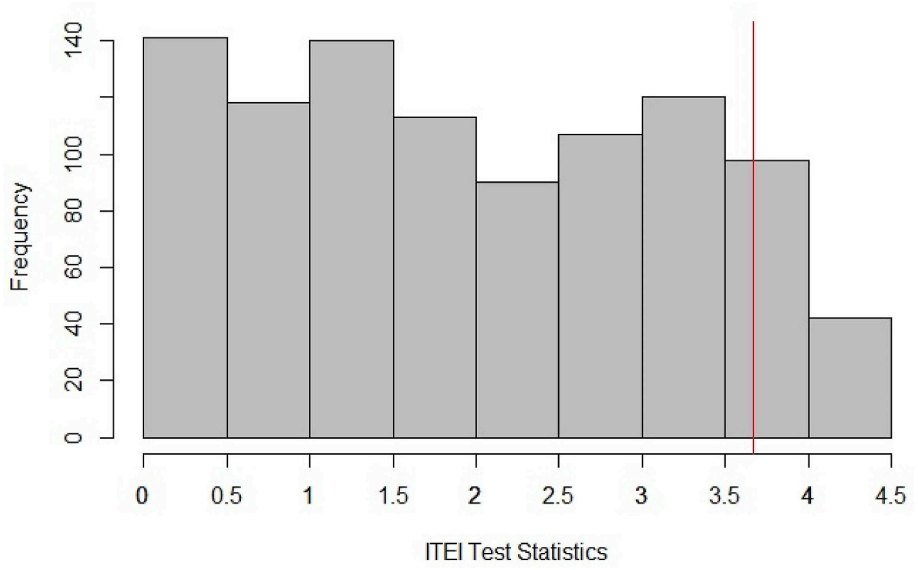

Fig. 2. Distribution of ITEIs given the restricted randomization scheme. The red line indicates the observed test statistic. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

remains the same.

Step 4: Obtaining the reference distribution and *p*-value

To obtain information about the significance of each data aspect, we can subsequently construct the reference distribution to locate the observed test statistic. In line with current quality standards for SCEDs (Kratochwill et al., 2010), we choose a restricted randomization scheme in which each phase must have a minimum of at least three data points. It should be noted that the original study did not include an element of randomization in the design. For a discussion on how randomization tests function without random assignment procedures as opposed to with random assignment procedures, we refer the interested reader to Ferron et al. (2003). Given the restricted randomization scheme, the three phase change moments from A1 to B1, B1 to A2, and A2 to B2 can now be shifted in such a way that each phase has at least three measurements while the order of the actually observed measurements remains intact (cf. Heyvaert & Onghena, 2014). As mentioned previously, there are 969 possibilities for shifting the three phase change moments given the restriction of at least three measurements per phase, one of which is the experiment as it was actually carried out. Calculating the test statistic(s) for the other 968 possibilities results in the reference distribution. A histogram of the reference distribution for ITEI is shown in Fig. 2.

The red vertical line indicates the observed test statistic of 3.67. All test statistics on the red line or to the right of it are as large as or even larger than the observed test statistic. There are 106 randomizations that would have led to a test statistic as extreme as 3.67 or more extreme. Accordingly, the *p*-value for immediacy equals $\frac{106}{969} = .11$. The same steps can be followed to calculate the *p*-values for level, trend, variability, and overlap. Because for consistency smaller test statistics indicate higher consistency, a slight modification is needed. For consistency, the *p*-value equals the proportion of test statistics that are as small as the observed test statistic or even smaller. The observed test statistic for each data aspect with corresponding *p*-values can be found in Table 2.

Table 2
Observed test statistic and p-value for each data aspect.

Data aspect	Observed test statistic	*p*-value
Level	2.74	.041
Trend	0.01	.947
Variability	15.30	.257
Overlap	1.00	.006
Immediacy	3.67	.109
Consistency	1.90	.466

4.2. Correcting the observed p-values for multiple testing

When performing several statistical tests an important consideration is how to control for potentially inflated Type I error rates. A Type I error –also referred to as false positive- occurs when a null hypothesis of no treatment effect is erroneously rejected. This is particularly relevant in SCEDs where researchers often assess multiple data aspects and dependent variables at the same time to reach an overall conclusion about a new treatment. Presenting each test result with a *p*-value of $\leq .05$ as statistically significant is a common misinterpretation, as it falls short of acknowledging that with an increasing number of tests, the risk of finding a false positive increases (Gordi & Khamis, 2004). The probability of at least one false positive occurring can be calculated as $1 - (1 - \alpha)^k$ with α being the significance level and k being the number of independent hypothesis tests (Sainani, 2009). Employing the usual threshold of $\alpha = .05$ for determining statistical significance, the risk of finding a false positive in our example with one dependent variable would already be $1 - (1 - .05)^6 = .26$. To better control the risk of finding false positives, several multiple testing corrections have been developed over the years.

Among the best known are the Bonferroni methods (see Armstrong [2014] for a review of the techniques). Bonferroni procedures control the Type I error rate familywise, meaning that the probability of falsely rejecting a single null hypotheses in the whole family of hypotheses should not be greater than the initially chosen α. In its simplest form the Bonferroni correction divides the α-level by the number of tests being run to adjust the significance threshold for the obtained *p*-values (Sainani, 2009). While this method indeed controls the Type I error rate, it greatly inflates Type II error rates, i.e. failing to reject a false null hypothesis (e.g., Armstrong, 2014; Garamszegi, 2006; Nakagawa, 2004; Sainani, 2009). Even adjustments to the classical Bonferroni method, such as step-up and step-down procedures, have received similar criticism as the original method (Verhoeven, Simonsen, & McIntyre, 2005).

Therefore, we consider a set of alternative techniques based on the false discovery rate (FDR). The FDR controls for a proportion of

erroneously rejected null hypotheses instead of trying to avoid a single Type I error at the cost of a loss in power (Benjamini & Hochberg, 1995, 2000). This is preferable for SCEDs where "the overall conclusion that the treatment is superior need not be erroneous even if some of the null hypotheses are falsely rejected" (Benjamini & Hochberg, 1995, p. 290). Specifically, we will adjust the *p*-values in Table 2 using three FDR-based techniques: the original FDR technique proposed by Benjamini and Hochberg (1995), herafter referred to as BH95; the modification of BH95 proposed by Benjamini and Hochberg (2000) to incorporate an estimator of the number of true null hypotheses, hereafter referred to as BH00; and finally a modification of BH95 proposed by Benjamini and Yekutieli (2001) for situations in which the dependency between tests is negative or unknown; hereafter referred to as BY01.

BH95 works as shown in Equation (12), in which k represents the number of null hypotheses tested and i represents the rank number of the ordered *p*-values.

$$P_{(i)} \leq \left(\frac{i}{k}\right)\alpha \tag{12}$$

First, the desired level of control α is chosen. Next, the *p*-values are arranged in ascending order where $P_{(i)}$ corresponds to the null hypothesis $H_{(i)}$. Starting with the largest *p*-value, each *p* is checked for the requirement shown in Equation (12) (Verhoeven et al., 2005). Let t be the largest i for which Equation (12) holds true, then reject all hypotheses H_1, H_2,..., H_t. BH95 then controls the FDR at α (Benjamini & Hochberg, 1995). BH95 has been shown to control the FDR for cases in which the test statistics are independent as well as positively correlated (Benjamini & Yekutieli, 2001). This method is available in the stats package in base R.

In BH00 Benjamini and Hochberg (2000) modified BH95 to incorporate an estimator $\hat{m}_0$ for the number of true null hypotheses as shown in Equation (13).

$$P_{(i)} \leq \left(\frac{i}{\hat{m}_0}\right)\alpha \tag{13}$$

"The unknown $\hat{m}_0$ is first estimated from the data, and then utilized by the adaptive procedure" (Benjamini & Hochberg, 2000, p. 62). Several R-packages such as "fdrtool" (Klaus & Strimmer, 2015) and "MuToss" (MuToss Coding Team, 2015) incorporate this procedure. After the estimator $\hat{m}_0$ has been determined, the procedure follows the same steps as BH95 and controls FDR at α. Verhoeven et al. (2005) have shown that BH00 can potentially offer better control of the Type II error rate while still controlling for Type I errors.

Finally, Benjamini and Yekutieli (2001) suggested BY01, which is a modification of BH95 for situations in which the dependency between tests is negative or unknown. Regarding BY01, Verhoeven et al. (2005) caution that "this modification is more conservative than the original procedure, and thus should be used only when made necessary by negative dependency among tests" (p. 644). This modification divides the level of control α by the sum of the inversed ranks of *p*-values as shown in Equation (14).

$$P_{(i)} \leq \left(\frac{i}{m}\right)\left(\frac{\alpha}{\sum_{i=1}^{m}\frac{1}{i}}\right) \tag{14}$$

The procedure then follows the same steps as BH95 and BH00 by checking each *p*-value sequentially against the criterion. BY01 is also available in the stats package in base R.

Table 3 presents the results per data aspect adjusted for each of the three multiple testing corrections. With the BH95 adjustment, the first and only data aspect that fulfills the criterion given in Equation (12) is overlap: $P_{(1)} = .006 < \left(\frac{1}{6}\right).05 = .008$. For overlap $i = 1$ because it was initially the lowest *p*-value. When adjusted with the BH95 correction, the initially marginally significant *p*-value for level (second lowest *p*-value) becomes insignificant: $P_{(2)} = .041 > \left(\frac{2}{6}\right).05 = .017$. The adjusted *p*-values shown in Table 3 for BH95 can be calculated using the stats package in base R. They are obtained by the formula $\frac{p_i}{i \div m}$. For the example data set, adjusting the *p*-values according to BH00 yields identical results as adjusting with BH95. This is the case because the estimated number of true null hypotheses ($\hat{m}_0$) is equal to six. We estimated $\hat{m}_0$ using the R-package "MuToss" (MuToss Coding Team, 2015). The adjusted *p*-values are then obtained in the same way as for BH95. Finally, when adjusted according to BY01, none of the data aspects are significant as even the lowest *p*-value (for overlap) does not meet the requirement shown in Equation (14): $P_{(1)} = .006 > \left(\frac{1}{6}\right)\left(\frac{.05}{2.45}\right) = .003$. The adjusted *p*-values can be calculated using the stats package in base R. They are obtained by the formula $\left(\frac{P_{(i)}}{i \div m}\right)\left(\sum_{i=1}^{m}\frac{1}{i}\right)$.

Table 3
P-value adjustments for three FDR techniques.

FDR technique	Data aspect	Adjusted *p*-value
BH95	Level	.124
	Trend	.947
	Variability	.385
	Overlap	.037*
	Immediacy	.219
	Consistency	.560
BH00	Level	.124
	Trend	.947
	Variability	.385
	Overlap	.037*
	Immediacy	.219
	Consistency	.560
BY01	Level	.303
	Trend	1
	Variability	.944
	Overlap	.091
	Immediacy	.536
	Consistency	1

*$p \leq .05$.

5. Discussion

In this article we presented a multiple randomization test procedure for analyzing level, trend, variability, overlap, immediacy, and consistency of data patterns in SCEDs as critical adjuncts to visual analysis. Firstly, we proposed quantifications for each data aspect. Subsequently, we used the calculated quantifications as test statistics in multiple randomization tests to obtain the *p*-value for each data aspect. Finally, we corrected the *p*-values for multiple testing with three corrections based on the FDR. We provide generic R-code in the online supplementary material to perform the presented analyses.

Using this approach, we re-analyzed the data from Feeney and Ylvisanker (2003). Our re-analysis was not intended to show flaws in the conclusions drawn by the authors. Most notably, the study was published in 2003 and should accordingly not be judged by today's standards for quality and visual analysis. It should also be noted that the study was published seven years before the first version of the What Works Clearinghouse guidelines was introduced. As mentioned throughout the manuscript, by today's standards for conducting and visually analyzing SCEDs, the authors may not have concluded that a functional relation exists (e.g., less than five data points per phase and trend in the therapeutic direction in A2). At the same time, the authors' clear and precise reporting of the study design, data collection, dependent variables, and data analysis facilitated the reconstruction of how the study was conducted. The authors chose their intervention based on a careful review of the existing literature and theories. Their analysis was reasonable and took into account the clinical significance of the results on the participant's life and environment as well as qualitative follow-up data for both participants. At the same time, it enabled us to point out a few general pitfalls of merely analyzing the data

visually paired with simple descriptive statistics without taking into account all data aspects recommended for visual analysis.

The multiple randomization tests approach enables us to refine the original conclusion of Feeney and Ylvisanker (2003) reached by visual analysis that the intervention "clearly" demonstrated that the frequency of challenging behavior was "dramatically" reduced. Firstly, we can assess the effect the intervention has on each data aspect via the obtained quantifications. For example, the intervention leads to a complete nonoverlap between each pair of adjacent AB phases (NAP = 1.00), but it does not lead to a change in trend ($f^2 = 0.01$) as the intervention phase continues the trend of the preceding baseline phase. The intervention leads on average to an immediate decrease of 3.67 occurrences of challenging behavior within the first three measurements after introducing the intervention. Taking into account all measurement occasions in each phase, the intervention decreases the frequency of challenging behavior on average by 2.74 instances. Finally, the intervention leads to an increase of variability in the dependent variable (VR = 15.3) and the consistency of data patterns in similar phases is low (CONDAP = 1.90). While the presented analysis helps refine the original conclusion, a word of caution is in order for blindly following the described procedure. Testing for all six data aspects at the same time comes with a decrease in power. For example, if the researchers do not expect a change in trend *a priori*, the hypotheses test for trend can be safely dropped from the overall analysis to increase power without loss of information. Similarly, it is important to assess the combination of *p*-values rather than each *p*-value separately. For example, if we find a significant *p*-value for level but not for trend, our confidence increases that there is actually a change in trend alone. At the same time, this change in level might be a continuation of a trend from the baseline phase into the intervention phase. If we find a significant *p*-value for both trend and level, this can indicate that an undesirable trend has been reversed so strongly that it also leads to a change in level. In both situations, the inclusion of visual analysis can help making sense of these interactions and strengthen the conclusions drawn from the experiment.

As mentioned before, the original experiment did not contain any element of randomization. The randomization tests in the present paper were therefore carried out under two assumptions. The first assumption was that the researchers had implemented a restricted randomization scheme that allows for at least three data points per phase. The second assumption was that the experiment as it was actually carried out was chosen at random from all possible arrangements permissible given the restricted randomization scheme. Ideally, an element of randomization is determined *a priori* so that one of the possible assignments can be chosen at random. This procedure increases the scientific validity of the experiment as well as statistical conclusion validity while decreasing threats to internal validity including history and maturation (Edgington, 1996; Ferron et al., 2003; Michiels & Onghena, 2018; Onghena & Edgington, 2005). Therefore, we strongly recommend incorporating an element of randomization in the study design. When applying randomization tests to data obtained from a study that did not include randomization in the design phase –as in the given example- Type I error control might deviate from the nominal level, especially in the presence of autocorrelation (Ferron et al., 2003).

In the example, the data aspects level and overlap were initially significant at the .05 level. When correcting the results for multiple testing, BH95 and BH00 yielded identical results leaving only overlap as significant. This indicates in no way that this will always happen. For the example data, BH95 and BH00 yielded identical results because the estimate of the number of true null hypotheses for BH00 was equal to six. With the BY01 correction none of the data aspects was found to be significant which strongly disagrees with a visual inspection of the graphed data. As Verhoeven et al. (2005) indicated, BY01 might be too conservative and should only be used when the tests are negatively correlated.

Combining visual analysis, effect size calculation, and statistical inference for SCEDs has the advantage that it makes the results interpretable to researchers familiar with different analytical techniques and coming from different backgrounds. We hope that this can help narrow the divide between applied researchers and methodologists and provide fruitful grounds for collaborations. In addition, the connection of these analytical techniques can help increase the scientific credibility of SCEDs outside the single-case community and make results from SCEDs more convincing. Finally, the presented approach addresses the broader issue concerning the replicability of research results, especially in psychology and educational sciences. Simmons et al. (2011) use the term *researcher degrees of freedom* for referring to the many choices applied researchers have to make before and during a study, e.g., how many data points to collect, which variables to analyze, or which methods for data analysis to use. The presented analysis tackles several of these issues as the described randomization procedure is determined *a priori*, fixing the number of measurements and method of data analysis beforehand. In addition, the method for correcting for multiple testing is determined *a priori* for the multiple randomization test procedure so that the inference criteria cannot be determined in an ad hoc manner (Wicherts et al., 2016).

5.1. Limitations and future research

To keep the demonstration of the multiple randomization tests approach manageable and easy to follow, we only used one dependent variable for one participant. This entails several limitations that offer potential avenues for future research. It would for example be interesting for future studies, preferably applied studies, to use this approach for all dependent variables and all participants in a given study. We anticipate that multiple testing corrections become even more relevant as the number of dependent variables and participants increases. Another potential avenue for future research could be the assessment of the performance of the various multiple testing corrections under different conditions as part of a simulation study. Factors of interest for such a simulation study might include the number of hypotheses tests carried out, the total number of measurement occasions within each data set and each phase, and the degree of autocorrelation. Also in the context of a simulation study, it might be interesting to manipulate the effect sizes for the different data aspects, for example a large effect for variability and a small effect for level. That way, the sensitivity of each test can be assessed under a variety of conditions. A third potential avenue for future research is the application of the presented approach beyond ABAB designs. A fourth avenue for future research might be the assessment of the presented approach in the context of meta-analyses. The presented approach might for example be integrated in a multilevel framework (cf. Onghena, Michiels, Jamshidi, Moeyaert, & Van den Noortgate, 2017). A final avenue for future research is the incorporation of the third demonstration of an effect when changing from the first intervention phase back to baseline measures. The change from intervention back to baseline is conceptually different from the changes from baseline to intervention (Tanious et al., 2019). Changes from baseline to intervention occur twice in an ABAB design whereas a change from intervention back to baseline occurs only once. The correct incorporation of the change from intervention back to baseline into one overarching statistical model is a major challenge for future research.

6. Conclusion

Following the growing consensus in the SCED literature that visual and statistical analyses are best used concurrently and addressing concerns about the replicability of psychological and educational research, we presented a multiple randomization tests approach for analyzing single-case ABAB data. The presented approach takes into account each data aspect routinely assessed by visual analysts: level, trend, variability, overlap, immediacy, and consistency of data patterns. The approach presented in this article adds to mere visual analysis by

giving information about the size of the effect and statistical significance of each data aspect while at the same time controlling for Type I error rates by employing multiple testing corrections. We hope to encourage applied researchers to use such an approach in the future to increase the transparency of reporting results from SCEDs. To facilitate implementation of the presented approach, we provide generic R-code.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.brat.2019.103414.

References

Armstrong, R. A. (2014). When to use Bonferroni correction. *Ophthalmic and Physiological Optics, 34*, 502–508.

Barton, E. E., Lloyd, B. P., Spriggs, A. D., & Gast, D. L. (2018). Visual analysis of graphic data. In J. R. Ledford, & D. L. Gast (Eds.). *Single case research methodlogy: Applications in special education and behavioral sciences* (pp. 179–214). (3rd ed.). New York: Routledge.

Beeson, P. M., & Robey, R. R. (2006). Evaluating single-subject treatment research: Lessons learned from the aphasia literature. *Neuropsychology Review, 16*, 161–169. https://doi.org/10.1007/s11065-006-9013-7.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B, 57*, 289–300.

Benjamini, Y., & Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *Journal of Educational and Behavioral Statistics, 25*, 60–83.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics, 29*, 1165–1188.

Beretvas, S. N., & Chung, H. (2008). A review of meta-analyses of single-subject experimental designs: Methodological issues and practice. *Evidence-Based Communication Assessment and Intervention, 2*, 129–141. https://doi.org/10.1080/17489530802446302.

Brossart, D. F., Parker, R. I., Olson, E. A., & Mahadevan, L. (2006). The relationship between visual analysis and five statistical analyses in a simple AB single-case research design. *Behavior Modification, 30*, 531–563. https://doi.org/10.1177/0145445503261167.

Bulté, I., & Onghena, P. (2013). The single-case data analysis package: Analysing single-case experiments with R software. *Journal of Modern Applied Statistical Methods, 12*, 450–478.

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.

Castro Sotos, A. E., Vanhoof, S., Van den Noortgate, W., & Onghena, P. (2007). Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review, 2*, 98–113. https://doi.org/10.1016/j.edurev.2007.04.001.

Center, B. A., Skiba, R. J., & Casey, A. (1985-86). A methodology for the quantiative synthesis of intra-subject design research. *Journal of Special Education, 19*, 387–400. https://doi.org/10.1177/002246698501900404.

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

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

Edgington, E. S. (1975a). Randomization tests for one-subject operant experiments. *The Journal of Psychology, 90*, 57–68.

Edgington, E. S. (1980). Overcoming obstacles to single-subject experimentation. *Journal of Educational Statistics, 5*, 261–267.

Edgington, E. S. (1996). Randomized single-subject experimental designs. *Behavior Research and Therapy, 34*, 567–574.

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

Fedorov, S. (2013). GetData Graph Digitizer [Computer software]. Retrieved from http://getdata-graph-digitizer.com/.

Feeney, T. J., & Ylvisanker, M. (2003). Context-sensitive behavioral supports for young children with TBI: Short-term effects and long-term outcome. *The Journal of Head Trauma Rehabilitation, 18*, 33–51.

Ferron, J., & Foster-Johnson, L. (1998). Analyzing single-case data with visually guided randomization tests. *Behavior Research Methods, Intruments & Computers, 30*, 698–706.

Ferron, J., Foster-Johnson, L., & Kromrey, J. D. (2003). The functioning of single-case randomization tests with and without random assignment. *The Journal of Experimental Education, 71*, 267–288.

Ferron, J., & Jones, P. K. (2006). Tests for the visual analysis of response-guided multiple-baseline data. *Journal of Experimental Education, 75*, 66–81. https://doi.org/10.3200/JEXE.75.1.66-81.

Ganz, J. B., & Ayres, K. M. (2018). Methodological standards in single-case experimental design: Raising the bar. *Research in Developmental Disabilities, 79*, 3–9. https://doi.org/10.1016/j.ridd.2018.03.003.

Garamszegi, L. Z. (2006). Comparing effect sizes across variables: Generalization without the need for Bonferroni correction. *Behavioral Ecology, 17*, 682–687.

Gordi, T., & Khamis, H. (2004). Simple solution to a common statistical problem: Interpreting multiple tests. *Clinical Therapeutics, 26*, 780–786.

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

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

Heyvaert, M., & Onghena, P. (2014). Analysis of single-case data: Randomization tests for measures of effect size. *Neuropsychological Rehabilitation, 24*, 507–527. https://doi.org/10.1080/09602011.2013.818564.

Heyvaert, M., Wendt, O., Van den Noortgate, W., & Onghena, P. (2015). Randomization and data-analysis items in quality standards for single-case experimental studies. *The Journal of Special Education, 49*, 146–156. https://doi.org/10.1177/0022466914525239.

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children, 71*, 165–179. https://doi.org/10.1177/001440290507100203.

Klaus, B., & Strimmer, K. (2015). Estimation of (local) false discovery rates and higher criticism. R package version 1.2.15. Available at: https://CRAN.R-project.org/package = fdrtool.

Kratochwill, T. R., & Levin, J. R. (2014). Meta- and statistical analysis of single-case intervention research data: Quantitative gifts and a wish list. *Journal of School Psychology, 52*, 231–235. https://doi.org/10.1016/j.jsp.2014.01.003.

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., et al. (2010). *Single-case designs technical documentation.* Retrieved from What Works Clearinghouse: http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf.

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., et al. (2013). Single-case intervention research design standards. *Remedial and Special Education, 34*, 26–38. https://doi.org/10.1177/0741932512452794.

Kromrey, J. D., & Foster-Johnson, L. (1996). Determining the efficacy of intervention: The use of effect sizes for data analysis in single-subject research. *The Journal of Experimental Education, 65*, 73–93. https://doi.org/10.1080/00220973.1996.9943464.

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

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. https://doi.org/10.1016/j.jsp.2012.05.001.

Maggin, D. M., Briesch, A. M., & Chafouleas, S. M. (2013). An application of the what Works Clearinghouse standards for evaluating single-subject research: Synthesis of the self-management literature base. *Remedial and Special Education, 34*, 44–59. https://doi.org/10.1177/0741932511435176.

Manolov, R. (2018). Linear trend in single-case visual and quantitative analyses. *Behavior Modification, 42*, 684–706. https://doi.org/10.1177/0145445517726301.

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

Matyas, T. A., & Greenwood, K. M. (1990). Visual analysis of single-case time series: Effects of variability, serial dependence, and magnitude of intervention effects. *Journal of Applied Behavior Analysis, 23*, 341–351. https://doi.org/10.1901/jaba.1990.23-341.

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. https://doi.org/10.3758/s13428-016-0714-4.

Michiels, B., & Onghena, P. (2018). Randomized single-case AB phase designs: Prospects and pitfalls. *Behavior Research Methods,* 1–23.

Moore, D. S., & McCabe, G. P. (2005). *Introduction to the practice of statistics* (5th ed.). New York: W.H. Freeman.

MuToss Coding Team (2015). *MuToss: Unified multiple testing procedures. R package version 0.1-10.* Retrieved from https://CRAN.R-project.org/package = mutoss.

Nakagawa, S. (2004). A farewell to Bonferroni: The problems of low statistical power and publication bias. *Behavioral Ecology, 15*, 1044–1045.

Olive, M. L., & Smith, B. W. (2005). Effect size calculations and single subject designs. *Educational Psychology, 25*, 313–324. https://doi.org/10.1080/0144341042000301238.

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

Onghena, P., & Edgington, E. S. (2005). Customization of pain treatments: Single-case design and analysis. *The Clinical Journal of Pain, 21*, 56–68. https://doi.org/10.1097/00002508-200501000-00007.

Onghena, P., Michiels, B., Jamshidi, L., Moeyaert, M., & Van den Noortgate, W. (2017). One by one: Accumulating evidence by using meta-analytical procedures for single-case experiments. *Brain Impairment, 19*, 33–58. https://doi.org/10.1017/BrImp.2017.2.

Ottenbacher, K. J. (1990). When is a picture woth a thousand p values? A comparison of visual and quantitative methods to analyze single subject data. *The Journal of Special Education, 23*, 436–449. https://doi.org/10.1177/002246699002300407.

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

Park, H.-S., Marascuilo, L., & Gaylord-Ross, R. (1990). Visual inspection and statistical analysis in single-case designs. *The Journal of Experimental Education, 58*, 311–320. https://doi.org/10.1080/00220973.1990.10806545.

Perdices, M., & Tate, R. L. (2009). Single-subject designs as a tool for evidence-based clinical practice: Are they unrecognised and undervalued? *Neuropsychological Rehabilitation, 19*, 904–927. https://doi.org/10.1080/09602010903040691.

Pustejovsky, J. E., Hedges, L. V., & Shadish, W. R. (2014). Design-comparable effect sizes in multiple baseline designs: A general modeling framework. *Journal of Educational and Behavioral Statistics, 39*, 368–393. https://doi.org/10.3102/1076998614547577.

Rvachew, S. (1988). Application of single subject randomization designs to communicative disorders research. *Human Communication Canada, 12*, 7–13.

Sainani, K. L. (2009). The problem of multiple testing. *PM&R, 1*, 1098–1103.

Scruggs, T. E., Mastropieri, M. A., & Casto, G. (1987). The quantitative synthesis of single-subject research: Methodology and validation. *RASE, 8*, 24–33. https://doi.org/10.1177/074193258700800206.

Shadish, W. R., Hedges, L. V., Horner, R. H., & Odom, S. L. (2015). *The role of between-case effect size in conducting, interpreting, and summarizing single-case research.* U.S. Department of Education, Institute of Education Sciences.

Shadish, W. R., Hedges, L. V., & Pustejovsky, J. E. (2014). Analysis and meta-analysis of single-case designs with a standardized mean difference statistic: A primer and applications. *Journal of School Psychology, 52*, 123–147. https://doi.org/10.1016/j.jsp.2013.11.005.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359–1366.

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.

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.

Tanious, R., De, T. K., Michiels, B., Van Den Noortgate, W., & Onghena, P. (2019). *Consistency in single-case ABAB phase designs: A systematic review.* Manuscript accepted.

Tanious, R., De, T. K., Michiels, B., Van den Noortgate, W., & Onghena, P. (2019). *Assessing consistency in single-case A-B-A-B phase designs. Behavior Modification.* Advance online publicationhttps://doi.org/10.1177/0145445519837726.

Tate, R. L., Perdices, M., Rosenkoetter, U., McDonald, S., Togher, L., Shadish, W. R., et al. (2016b). The single-case reporting guideline in BEhavioural interventions (SCRIBE) 2016: Explanation and elaboration. *Archives of Scientific Psychology, 4*, 1–9. https://doi.org/10.1037/arc0000026.

Tate, R. L., Perdices, M., Rosenkoetter, U., Shadish, W. R., Vohra, S., Barlow, D. H., et al. (2016a). The Single-Case Reporting guideline in BEhavioural interventions (SCRIBE) 2016 statement. *Aphasiology, 30*, 862–876. https://doi.org/10.1080/02687038.2016.1178022.

U.S. Department of Education, Institute of Education Sciences (2016, March). *What works Clearinghouse.* Retrieved from https://ies.ed.gov/ncee/wwc/Docs/ReferenceResources/wwc_srg_scd_instructions_s3_v2.pdf.

Verhoeven, K. J., Simonsen, K. L., & McIntyre, L. M. (2005). Implementing false discovery rate control: Increasing your power. *Oikos, 108*, 643–647.

Weaver, E. S., & Lloyd, B. P. (2018). Randomization tests for single case designs with rapidly alternating conditions: An analysis of p-values from published experiments. *Perspectives on Behavior Science,* 1–29. https://doi.org/10.1007/s40614-018-0165-6.

Welch, B. L. (1937). On the z-test in randomized blocks and Latin squares. *Biometrika, 29*, 21–52.

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., van Aert, R. C., & Van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology, 7*, 1–12. https://doi.org/10.3389/fpsyg.2016.01832.

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594–604. https://doi.org/10.1037/0003-066X.54.8.594.