Behavior Research Methods (2019) 51:1145–1160
https://doi.org/10.3758/s13428-018-1044-5

Nonparametric meta-analysis for single-case research: Confidence intervals for combined effect sizes

Bart Michiels[1] · Patrick Onghena[1]

Published online: 16 April 2018

Abstract

In this article we present a nonparametric technique for meta-analyzing randomized single-case experiments by using inverted randomization tests to calculate nonparametric confidence intervals for combined effect sizes (CICES). Over the years, several proposals for single-case meta-analysis have been made, but most of these proposals assume either specific population characteristics (e.g., heterogeneity of variances or normality) or independent observations. However, such assumptions are seldom plausible in single-case research. The CICES technique does not require such assumptions, but only assumes that the combined effect size of multiple randomized single-case experiments can be modeled as a constant difference in the phase means. CICES can be used to synthesize the results from various single-case alternation designs, single-case phase designs, or a combination of the two. Furthermore, the technique can be used with different standardized or unstandardized effect size measures. In this article, we explain the rationale behind the CICES technique and provide illustrations with empirical as well as hypothetical datasets. In addition, we discuss the strengths and weaknesses of this technique and offer some possibilities for future research. We have implemented the CICES technique for single-case meta-analysis in a freely available R function.

Keywords Single-case experiments · Meta-analysis · Effect size · Confidence intervals · Hypothesis testing · Nonparametric statistics · Randomization tests

Meta-analysis is one of the primary methods for identifying effective treatments in education, clinical psychology, medical science, and many other fields, as it can lead to more reliable conclusions about specific treatments through synthesizing the results of many individual studies (Shadish, Hedges, & Pustejovsky, 2014). Single-case experiments (SCEs) are regarded as a group of experimental designs that allow for strong causal inferences and that therefore should be included in meta-analyses to inform evidence-based decision making (Shadish & Rindskopf, 2007). SCEs are designed experiments that are suitable to assess the efficacy of a treatment for a single case. In such experiments, repeated measurements are recorded for this case on at least one dependent variable under different levels (i.e., treatments) of one or more independent variables (Barlow, Nock, & Hersen, 2009; Gast & Ledford, 2014; Kazdin, 2011; Onghena, 2005). Note that the "single case" can refer to a unit at various levels of aggregation, such as a person, a family, a classroom, or a school.

The last decade has seen a growing interest in SCE meta-analysis (Maggin, O'Keeffe, & Johnson, 2011; Shadish & Rindskopf, 2007) and various meta-analytic techniques have been proposed. The most important statistical proposals have been the calculation of various measures of effect size (ES; e.g., Burns, 2012; L. G. Hedges, Pustejovsky, & Shadish, 2012, 2013; Heyvaert et al., 2017; Kratochwill & Levin, 2014; Parker, Vannest, & Davis, 2011; Scruggs & Mastropieri, 2013) and the use of multilevel models (e.g., Ferron, Bell, Hess, Rendina-Gobioff, & Hibbard, 2009; Ferron, Farmer, & Owens, 2010; Moeyaert, Ugille, Ferron, Beretvas, & Van den Noortgate, 2014; Nugent, 1996; Shadish & Rindskopf, 2007; Van den Noortgate & Onghena, 2003). In the following paragraphs we will briefly discuss both categories of meta-analytic techniques.

Calculating effect size (ES) measures is regarded as extremely important for reporting scientific results (Cohen, 1990, 1994; Kirk, 1996). Moreover, major scientific organizations highly recommend that measures of ES and

✉ Bart Michiels
Bart.Michiels@ppw.kuleuven.be

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

confidence intervals are reported in addition to *p* values from statistical tests (American Psychological Association, 1994, 2001, 2010; Wasserstein & Lazar, 2016; Wilkinson & the Task Force on Statistical Inference, 1999). Major advantages of calculating standardized ES measures is that they allow the comparison of ESs from different studies and that they can be used for meta-analyzing the results of multiple individual studies (Shadish, 2014).

Hedges, Pustejovsky, and Shadish (2012, 2013) have proposed a standardized *d* statistic as an ES measure for single-case meta-analysis, comparable to the *d* statistic used in between-subjects designs. An advantage of this ES measure is that the proposed *d* statistic has a formal statistical development that makes significance testing and the construction of confidence intervals possible. A disadvantage of this method is that several distributional assumptions have to be made in order to enable statistical inference such as homogeneity of within-case errors and between-case variation, an AR1 autocorrelation structure for the within-case errors and normality for the between-case distribution. Another limitation is that the *d* statistic can only be used for AB^k phase designs or multiple-baseline designs but not for single-case designs that are based on rapid treatment alternation such as an alternating treatments design or a randomized block design.

Some authors have proposed the use of nonoverlap statistics for meta-analytic purposes in single-case research (Parker et al., 2011; Scruggs & Mastropieri, 2013). This approach entails calculating a specific nonoverlap statistic for each individual study included in the meta-analysis and then averaging the individual values to obtain an average ES. An advantage of this approach is that most nonoverlap statistics are easy to calculate and interpret. A disadvantage is that most nonoverlap measures are not based on formal statistical theory, which makes parametric significance testing impossible. A small group of nonoverlap measures are formally developed (e.g., the nonoverlap of all pairs: Parker & Vannest, 2009; the percentage of all nonoverlapping data: Parker, Hagan-Burke, & Vannest, 2007), but the statistical tests that can be used to calculate *p* values for these measures (e.g., the Mann–Whitney *U* test for the nonoverlap of all pairs, and the chi-square test for the percentage of all nonoverlapping data) assume that the data consist of independent observations, which is often a questionable assumption in single-case research (Dugard, 2014; Shadish & Sullivan, 2011; Solomon, 2014).

Multilevel models have been proposed to synthesize the results of multiple individual SCEs (e.g., Ferron et al., 2009; Ferron et al., 2010; Moeyaert et al., 2014; Nugent, 1996; Shadish & Rindskopf, 2007; Van den Noortgate & Onghena, 2003). Multilevel models allow to estimate various parameters such as case-specific intercepts and treatments effects, the average treatment effect over all included SCEs as well as the within- and between-case variance of the treatment effect. Multilevel models are most frequently used to analyze data from multiple-baseline designs but can also be used to integrate multiple single-case phase designs on a particular topic. An advantage of this method is that it is a very flexible way to model single-case data patterns. A disadvantage is that significance testing and the calculation of standard errors of treatment effects in this approach are done using *t* procedures. These procedures rely on classical distributional assumptions and an assumption of random sampling, assumptions that are often not fulfilled in single-case research (Dugard, 2014). Furthermore multilevel models generally require quite large sample sizes for accurate standard error estimation of all model parameters (Maas & Hox, 2004) whereas single-case research often features rather small datasets.

Rindskopf (2014) has proposed a Bayesian variant of multilevel models for synthesizing single-case research. The author argues that the main advantage of Bayesian parameter estimation in multilevel models is that it does not require large sample sizes whereas maximum likelihood parameter estimation (i.e., the frequentist approach) does require large sample sizes for accurate parameter estimation (Rindskopf, 2014). Although the Bayesian approach is better at applying the multilevel model to small datasets, it is still required to make distributional assumptions for the outcome variable (the author assumes a binomially distributed outcome variable).

Meta-analyzing single-case experiments using randomization tests

The goal of this article is to propose a statistical technique that can be used to make statistical inferences and generate confidence intervals for the average treatment effect of multiple randomized single-case experiments, without resorting to distributional assumptions about the data or to an assumption of random sampling. Note that we use the term *single-case meta-analysis* to refer to data synthesis from multiple SCEs in general, regardless of whether they are replicated SCEs from the same researchers or various SCEs that were conducted at different times and by different researchers. Although data synthesis is conceptually different in these two situations, there are no statistical implications for the technique that we will propose. The technique that we will introduce is based on the random-assignment model that forms the foundation of the randomization test (RT). The RT has been proposed as an appropriate statistical test to evaluate treatment effects in randomized SCEs (e.g., Bulté & Onghena, 2008; Edgington, 1967; Ferron & Levin, 2014; Heyvaert & Onghena, 2014; Levin, Ferron, & Kratochwill, 2012; Levin, Marascuilo, & Hubert, 1978; Onghena, 1992; Onghena & Edgington, 1994, 2005). Therefore, from the outset we want to emphasize that this technique is intended for single-case designs that incorporate some form of random assignment. In the Discussion

section we will explain the implications for the proposed technique when it is used with nonrandomized designs.

Four steps have to be taken in order to use an RT in the evaluation of an individual SCE. First, and prior to executing the actual experiment, all *permissible assignments* for the chosen experimental design are listed. A permissible assignment is a random assignment of measurement occasions to treatment conditions that conforms to the restrictions imposed by the chosen randomization scheme. Second, one permissible assignment is randomly selected as the assignment for the actual experiment. Third, a test statistic that is adequate to answer the research question is chosen. RTs can provide one-sided or two-sided p values depending on whether the chosen test statistic is sensitive to the direction of the alternative hypothesis. For an RT with a two-sided p value, a nondirectional test statistic (e.g., an absolute mean difference) has to be used. Fourth, and after the data are collected, the randomization distribution is constructed by calculating the value of the test statistic for all permissible assignments, conditional on the observed data and their temporal ordering. This randomization distribution is used as the reference distribution to determine the statistical significance of the observed test statistic: The two-sided p value of an RT is calculated as the proportion of test statistics in the randomization distribution that are at least as extreme as the observed test statistic. Depending on the chosen significance level, one then either rejects or accepts the null hypothesis on the basis of the p value. See Heyvaert and Onghena (2014) for more details about this randomization test procedure.

It is also possible to use the RT to assess the statistical significance of the combined treatment effect of multiple individual SCEs. This can be done by using *combined assignments* (i.e., assignments that comprise the individual assignments of multiple individual SCEs) in the RT. We will illustrate this concept with an example.

Consider the following two hypothetical SCEs: an AB design and an alternating treatments design (ATD), both featuring ten measurement occasions. Each randomized SCE is associated with a collection of permissible assignments. This collection is determined by the type of single-case experimental design that is used and the number of measurement occasions in the experiment (Onghena, 2005). For illustration purposes, Table 1 displays a nonexhaustive set of three permissible assignments for each of the two hypothetical SCEs.

Suppose we select AAABBBBBBB and ABAABABBAB as the random assignments to execute the AB design and the ATD, respectively. Table 2 displays the hypothetical data for both designs alongside the condition labels of the employed random assignments.

One can conduct individual RTs for each experiment separately. For a single experiment, this can be done by first calculating the test statistics for all permissible assignments in order to obtain the randomization distribution. Next, the p value can be calculated by determining the proportion of test statistic values in the randomization distribution that are equal to or exceed the observed value of the test statistic. Table 3 displays the randomization distribution for each of the two SCEs using the absolute mean difference between the A observations and the B observations as the test statistic.

Table 1 A selection of three permissible assignments (PA) for a single-case AB design and an alternating treatments design (ATD)

Design	MO	1	2	3	4	5	6	7	8	9	10
AB	PA1	A	A	A	B	B	B	B	B	B	B
	PA2	A	A	A	A	A	B	B	B	B	B
	PA3	A	A	B	B	B	B	B	B	B	B
ATD	PA1	A	B	A	A	B	A	B	B	A	B
	PA2	B	A	B	B	A	B	A	A	B	A
	PA3	B	B	A	A	B	A	A	B	A	B

MO = measurement occasion

From Table 3 we can see that the p value for the AB design is 1/3 or .33, and the p value for the ATD is 2/3 or .67.

In the combined-assignment RT, the permissible assignments and their respective data values from both SCEs are integrated into a combined assignment. In this approach we calculate the chosen test statistic for each individual SCE and then average the resulting values into a single test statistic value for the combined assignment. Calculating the randomization distribution for the combined-assignment RT then consists of constructing the set of all possible combined assignments for the two SCEs and calculating the selected test statistic for each of them. To construct a combined assignment for these two SCEs, one simply selects a permissible assignment from each SCE and combines them into a single assignment. In our example, there are three permissible assignments per SCE, and thus nine different combined assignments. Table 4 displays the value of the absolute mean difference of the A observations and the B observations for each of the combined assignments.

For the combined-assignment RT, the p value is 1/9 or .1111. Note that this p value is considerably smaller than the p values of the individual RTs. As such, the combined-assignment RT can achieve higher statistical power than RTs that are applied to separate SCEs.

One can construct combined assignments for any number of individual SCEs using any combination of single-case randomization schemes. In general, the number of permissible combined assignments equals

$$\prod_{i=1}^{n} k_i$$

with n being the number of individual SCEs and k_i being the number of permissible individual assignments of SCE i. It is important to emphasize that the combined-assignment RT is

Table 2 Hypothetical datasets for a single-case AB design and an alternating treatments design (ATD), along with their respective random assignments

Design	AB										ATD									
RA	A	A	A	B	B	B	B	B	B	B	A	B	A	A	B	A	B	B	A	B
DV	4	5	4	7	8	8	9	8	8	7	3	6	4	4	6	4	7	8	5	7

RA = random assignment, DV = data values

perfectly valid if the randomization possibilities of each individual study are respected within each combined assignment.

Note that the number of permissible combined assignments increases exponentially with the number of individual studies included in the RT. For this reason we will use a random sample of all permissible combined assignments to execute the RT (also known as a *Monte Carlo RT*; e.g., Dwass, 1957; Hope, 1968). Edgington and Onghena (2007) demonstrated that a Monte Carlo RT is a valid test in its own right. In addition, the accuracy of the Monte Carlo RT can be increased to any desired level by increasing the size of the random sample. Edgington (1969) showed that an efficient Monte Carlo RT can be carried out with as few as 1,000 random assignments.

The combined-assignment RT can be used to compute a nonparametric *p* value for the combined ES across all included individual studies. The main drawback of reporting *p* values is that they do not provide information regarding the size of the effect (Cumming, 2014). Furthermore, there is currently a widespread consensus that ESs and confidence intervals should be reported in addition to *p* values (American Psychological Association, 1994, 2001, 2010; Wasserstein & Lazar, 2016; Wilkinson & the Task Force on Statistical Inference, 1999). In the next section we will describe how one can construct a nonparametric confidence interval for the combined ES of multiple studies using the combined-assignment RT. We will further refer to this technique as *confidence intervals for combined effect sizes* (CICES).

Constructing confidence intervals for combined effect sizes

The CICES technique is based on the principle of *hypothesis test inversion* (HTI; Garthwaite, 2005; Tritchler, 1984). This principle uses the equivalence between a $100(1-\alpha)\%$ two-sided confidence interval and a two-sided hypothesis test at significance level α (Neyman, 1937). Both are equivalent in the sense that for a certain test statistic θ, the $100(1-\alpha)\%$ two-sided confidence interval contains all point null values of θ that cannot be rejected by a two-sided hypothesis test at significance level α. HTI can be used with either parametric hypothesis tests or nonparametric hypothesis tests such as RTs. Michiels, Heyvaert, Meulders, and Onghena (2017) already showed that nonparametric confidence intervals for single-case ESs can be constructed from the RT using HTI.

To construct nonparametric CICES of multiple SCEs, it is necessary to assume an effect function that models the nature of the combined treatment effect in the data. For this purpose, the "unit-treatment additivity model" is one particular model that is most popular and well-studied within nonparametric statistics (e.g., Cox & Reid, 2000; Hinkelmann & Kempthorne, 2005, 2008, 2012; Lehman, 1959; Welch & Gutierrez, 1988). For a single experiment with two conditions, this model assumes that the experimental condition has a constant additive effect (denoted by Δ) on the scores of the outcome variable in the control condition. A simple illustration will clarify the rationale of the model. Consider an AB design with five measurement occasions in the baseline phase (A) and five measurement occasions in the treatment (phase). Table 5 illustrates how the unit-treatment additivity model expresses the observed scores.

The model assumes a *null score* for each measurement occasion (i.e., the score that would be observed for that measurement occasion when the null hypothesis of no treatment

Table 3 Test statistic values of all permissible assignments for the AB design and the alternating treatments design (ATD)

AB Design	$\lvert\overline{A}-\overline{B}\rvert$	ATD Design	$\lvert\overline{A}-\overline{B}\rvert$
AAABBBBBBB	**3.5238**	**ABAABABBAB**	**2.8**
AAAAABBBBB	2.4	BABBABAABA	2.8
AABBBBBBBB	2.875	BBAABAABAB	1.2

The employed random assignment and the corresponding observed test statistic value are marked in bold

Table 4 Test statistic values of all combined assignments for the AB design and the alternating treatments design

Combined Assignment	$\lvert\overline{A}-\overline{B}\rvert$
AAABBBBBBBABAABABBAB	**3.2917**
AAABBBBBBBBABBABAABA	0.3750
AAABBBBBBBBBAABAABAB	2.4583
AAAAABBBBBABAABABBAB	2.6
AAAAABBBBBBABBABAABA	0.2
AAAAABBBBBBBAABAABAB	1.8
AABBBBBBBBABAABABBAB	3.011
AABBBBBBBBBABBABAABA	0.0659
AABBBBBBBBBBAABAABAB	2.1319

The employed combined random assignment and the corresponding observed test statistic value are marked in bold

Table 5 Illustration of the way in which the unit-treatment additivity model expresses the observed scores of a single-case AB phase design

MO	1	2	3	4	5	6	7	8	9	10
CL	A	A	A	A	A	B	B	B	B	B
DV	$\mathbf{X_1}$	$\mathbf{X_2}$	$\mathbf{X_3}$	$\mathbf{X_4}$	$\mathbf{X_5}$	$\mathbf{X_6 + \Delta}$	$\mathbf{X_7 + \Delta}$	$\mathbf{X_8 + \Delta}$	$\mathbf{X_9 + \Delta}$	$\mathbf{X_{10} + \Delta}$

MO = measurement occasion, CL = condition labels, DV = data values

effect is true), along with a constant additive treatment effect (Δ) that is added to the null scores of the measurement occasions in the treatment phase. In other words, the model expresses the difference between the observed scores (denoted by Y) and the null scores (denoted by X) as

$Y_i = X_i + D_i\Delta$,with $i = 1 \ldots n$ measurement occasions, $D_i = 0$ for the measurement occasions in the A phase, and $D_i = 1$ for the measurement occasions in the B phase. Consequently, the null hypothesis (H_0) under this model can be written as

$H_0 : Y_i = X_i$.

Note that the adequacy of the unit-treatment additivity model depends on the specific data pattern analyzed and that alternative models can be used as an effect size function in CICES. We will come back to this issue in the Discussion section. In this article, we will use a constant Δ value for the combined treatment effect across all SCEs included in the meta-analysis. Constructing the nonparametric CICES then boils down to performing the combined-assignment RT for a range of hypothesized Δ values and retaining the nonrejected Δ values as part of the interval. The test statistic that we will use as the combined ES across all included SCEs in this illustration is the pooled average absolute mean difference between all A observations and all B observations (we will denote this test statistic as θ). However, the CICES technique can be used with different types of test statistics. The R function we have written for the CICES technique currently supports three test statistics: on the one hand the absolute mean phase difference as an unstandardized measure and on the other hand the Hedges, Pustejovsky, and Shadish (2012) d statistic and the Hedges's (1981) g statistic with correction for small sample size bias as standardized measures. In the CICES method, the selected ES measure is always calculated for each included SCE separately and then averaged across all the individual studies to produce the combined ES.

We will now discuss the steps that need to be taken to construct a nonparametric confidence interval for θ. First, we calculate the observed value of θ (θ_{obs}). Second, we use the unit-treatment additivity model (see above) to conduct a combined-assignment RT for different values of Δ. The null hypothesis of the RT can be rewritten as

$$Y_{iA} : Y_i = X_i$$

$Y_{iB} : Y_i = X_i + \Delta$,with Y_{iA} being the observed scores in the A phase and Y_{iB} being the observed scores in the B phase. For each Δ value, we test the null hypothesis that $Y_{iB} - Y_{iA} = \Delta$ against the alternative hypothesis that $Y_{iB} - Y_{iA} \neq \Delta$. To make the data correspond to the tested null hypothesis, we subtract Δ from the B scores of the observed data and perform the RT. Third, the randomization distribution for θ is constructed by selecting a random sample (of size j) of combined assignments and calculating the selected test statistic for all j combined assignments. Fourth, we use the randomization distribution to derive the p value. A p value equal to or smaller than the selected α level indicates that the tested Δ value is part of the $100(1-\alpha)\%$ confidence interval of θ. In practice we will use a computer algorithm that uses θ_{obs} as the initial value for Δ to perform the first RT and then incrementally increases the size of Δ according to a prespecified step size until the boundaries of the confidence interval are reached. The most computationally efficient way to calculate the confidence interval is to start with a relatively large step size to get a rough estimate of the boundaries quickly and then starting the algorithm again from the last Δ that yielded a p value smaller than the significance level but now with a step size that is ten times smaller. This iterative stepwise procedure can be used to calculate the confidence interval boundaries to any number of decimal places.

We will provide three illustrations of the CICES method. The first illustration uses empirical data from replicated ABAB designs and the second illustration uses empirical data from a multiple-baseline design (MBD). The third illustration uses hypothetical data to show how the results from SCEs using different types of experimental designs can be synthesized.

Illustration 1: Lambert et al. (2006)

As a first illustration of the CICES method, we use data from a study by Lambert, Cartledge, Heward, and Lo (2006), which evaluated the effects of response cards on the disruptive behavior of students in two urban fourth-grade classrooms. The study utilized an ABAB design, replicated over nine students, in which two conditions (A = "single-student responding" and B = "write-on response cards") were compared. The dependent variable was defined as the number of times a disruptive behavior was recorded by the observers during the experiment. See Lambert et al. for more details about this study and for their definition of "disruptive behavior."

This dataset was also reanalyzed in an article by Shadish, Hedges, and Pustejovsky (2014), using Hedges et al.'s (2012)

d statistic. This provides an opportunity to compare the confidence interval that Shadish et al. reported to the confidence interval of the CICES technique for the same dataset.

Figure 1 displays the Lambert et al. (2006) data. Note that the dataset contains some missing values. Shadish et al. (2014) dealt with this issue by omitting the measurement occasions with missing values from the analysis. CICES handles missing data by keeping the measurement occasions with missing values in the analysis (to keep the random assignment of measurement occasions to treatment conditions intact) but calculating the test statistic from only the available data points. Edgington and Onghena (2007) showed that the presence of missing data does not compromise the validity of the RT, provided that the occurrence of missing data is caused by a random factor and not a structural one. Because CICES is based on the RT rationale, the previous statement also holds for CICES. Note that in this sense CICES can also be used to synthesize the results from multiple-probe designs that feature probe-based measurement of the outcome variable rather than continuous measurement (Horner & Baer, 1978).

CICES, based on RT inversion, needs a randomization model. In this respect, it is important to notice that Lambert et al. (2006) do not mention that the intervention points for their ABAB design were randomized a priori, and from examining their graphs it is very unlikely that they were. For illustration purposes, we will assume that their SCE is randomized. However, it is important to note that the validity of the CICES inference cannot be guaranteed for nonrandomized SCEs, since the random-assignment assumption is not fulfilled. Hence, CICES results for nonrandomized SCES should be interpreted with caution. We will come back to this issue in the Discussion section. Following the recommendations by Kratochwill et al. (2010), that a phase should contain at least three measurement occasions, we will assume that the minimally required length of each phase in the ABAB design was three measurement occasions. The research question we want to answer is whether there is an average treatment effect of the response card treatment on the disruptive behavior of these nine students. To answer this question, we will calculate a nonparametric 95% CICES using Hedges et al.'s (2012) *d* statistic. A few steps need to be taken to use the CICES technique for the analysis of this dataset.

First, we calculate Hedges et al.'s (2012) *d* statistic for the Lambert et al. (2006) data. The observed value of the test statistic (θ_{obs}) is 2.51. We select 2.51 as the first Δ value, subtract Δ from all the B phase observations across all included studies and execute the combined-assignment RT. Second, we construct the set of combined assignments. The number of permissible assignments in a single-case phase design is given by the following formula (Onghena, 1992):

$\left(\frac{(N-n)(k+1)+k}{k}\right)$, with N being the number of measurement occasions, n being the minimum amount of measurement occasions in each phase and k being the number of phase changes in the design. Table 6 displays the number of possible assignments for each individual SCE of the Lambert et al. data.

To obtain the total number of possible combined assignments, we multiply all numbers of possible assignments for each individual study. This value runs into the billions, and even with the speed of today's computers, it would be unfeasible to calculate the test statistic for every possible combined assignment. For this reason, we select only a random sample (of size j) of the combined assignments for which we will calculate the test statistic, and perform a so-called Monte Carlo RT. For this example, we will use a random sample of size 5,000. Fourth, we construct the randomization distribution by calculating the test statistics for all 5,000 combined assignments. Fifth, we derive the p value from the randomization distribution. If we take the observed value of the test statistic as the first Δ value, then the p value takes on the maximal value of 1, which is larger than any conventional significance level and indicates that this value (2.51 in the example) is in the confidence interval. Next, we select a larger Δ value and repeat all of the previous steps. We keep selecting a larger Δ value until we reach the end of the confidence interval to the desired number of decimal places. Using the search algorithm implemented in the R function we have developed (available from https://ppw.kuleuven.be/home/english/research/mesrg/appletsandsoftware), the resulting 95% confidence interval at a precision of two decimal places is [1.47 ; 3.55]. The fact that the value of zero is not included in the confidence interval indicates that the response card treatment had a statistically significant effect on the number of disruptive behaviors across all the included students, given a 5% significance level. The corresponding two-sided p value of the combined-assignment RT is .0002, which also indicates a statistically significant treatment effect.

The confidence interval can be interpreted as follows: If the authors were to repeat their entire experiment a large number of times and subsequently analyze the resulting data, the observed value of Shadish et al.'s (2014) *d* statistic (2.51) would be contained in the nonparametric confidence interval in 95% of the repetitions (Moore, McCabe, & Craig, 2014). In addition, the interpretation of this nonparametric confidence interval is valid without having to make specific distributional assumptions, such as the equality of variances or normality of the data, or an assumption of random sampling.

Hedges et al.'s (2012) *d* statistic has a formal statistical development that makes it possible to estimate the variance of this measure and consequently to construct a confidence interval for it. Shadish et al. (2014) reported a 95% confidence interval of [2.12 ; 2.91] for the Lambert et al. (2006) data. We will return to the difference between the confidence interval reported by Shadish et al. (2014) and the confidence interval from the CICES technique in the Discussion section.

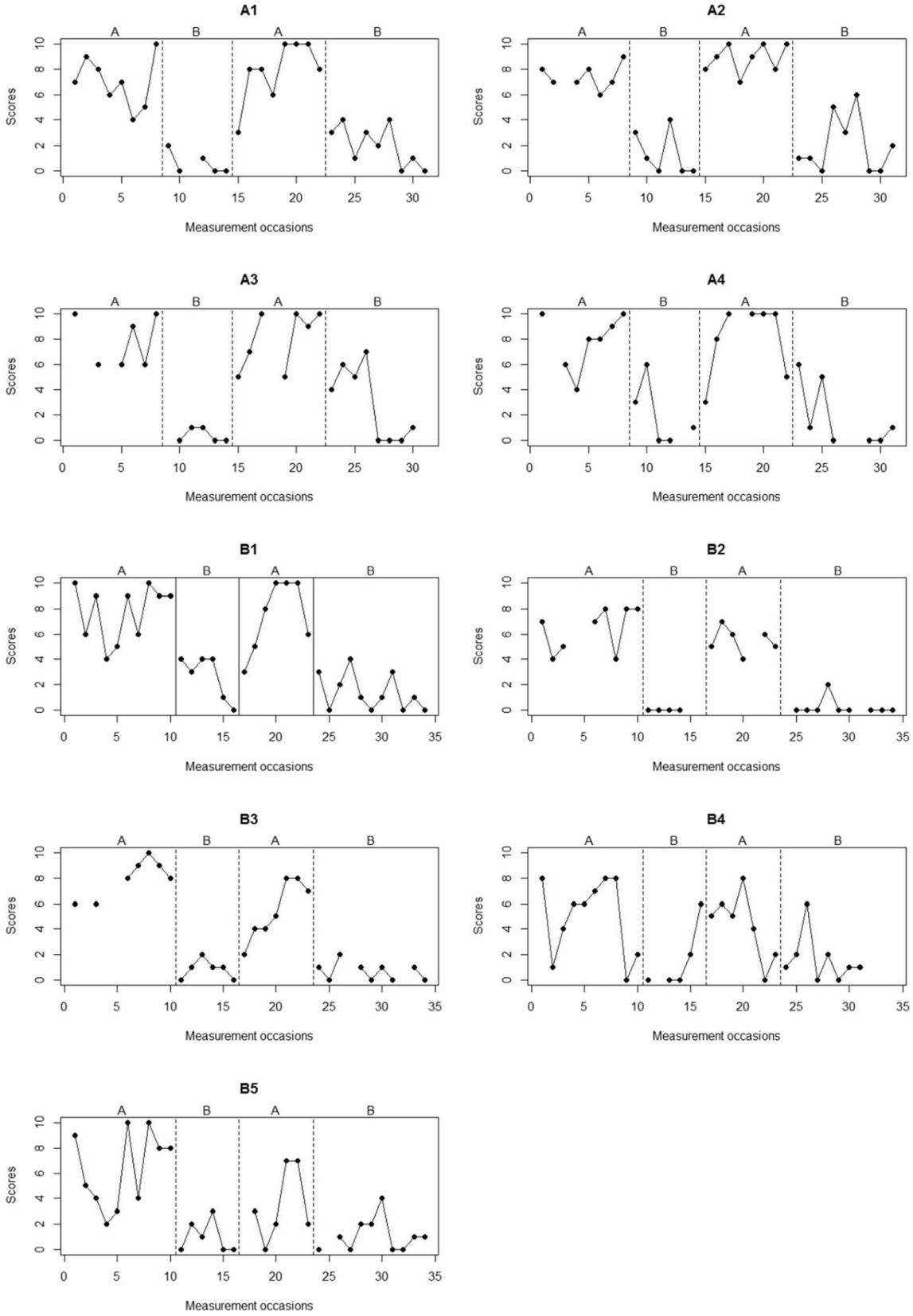

Fig. 1 Lambert et al. (2006) data for the "disruptive behavior" dependent variable for four (A1 to A4) and five (B1 to B5) students in two separate classrooms. Note that there are some gaps in the graphs due to missing data. The phases marked by "A" are the baseline phases, whereas the phases marked by "B" are the treatment phases

Table 6 Number of possible assignments (PA) for each individual single-case experiment of the Lambert et al. (2006) data

SCE	*N*	PA
1	31	1,540
2	31	1,540
3	30	1,330
4	31	1,540
5	34	2,300
6	34	2,300
7	34	2,300
8	31	1,540
9	34	2,300

Illustration 2: Multiple-baseline design

As a second illustration of the CICES technique, we use data from a study by Boersma, Linton, Overmeer, Jansson, Vlaeyen, and de Jong (2004). This study evaluated a graded-exposure in vivo treatment (GEXP; Vlaeyen, de Jong, Geilen, Heuts, & van Breukelen, 2001) for lowering fear–avoidance behavior and enhancing function in patients with chronic back pain and a high fear of movement (re)injury. The study used a multiple-baseline design with six participants. Each individual SCE consisted of a baseline phase followed by a treatment phase followed by a posttest phase. However, the staggering of intervention points only pertained to the start of the treatment phase, not to the start of the posttest phase. For this reason, we will limit our reanalysis to the data from the baseline phase and the treatment phase. The Boersma et al. dataset contains some missing data, which we will omit from the analysis. "Fear–avoidance" was the primary dependent variable. Fear–avoidance ratings (on a scale from 1 to 10) were recorded daily throughout a ten-week time period for all participants. The GEXP treatment was initiated at different time points for each participant, with a one-week interval between participants. Figure 2 displays the Boersma et al. data for the first baseline phase and the treatment phase. See Boersma et al. for graphs that include the data from the posttest.

The research question we want to answer is whether the GEXP treatment had an effect on fear and avoidance. To demonstrate the CICES technique with an unstandardized ES measure, we now choose the mean score differences between the baseline phase and the treatment phase, averaged over each participant, as the test statistic. The observed value of this test statistic is 2.38. Using a precision of two decimal places, the 95% confidence interval for the average treatment effect across all six participants is [– 0.93 ; 5.70]. The 95% confidence interval indicates that the GEXP treatment does not have a significant effect on fear–avoidance beliefs. The corresponding two-sided *p* value of the combined-assignment RT is .6456. This result contradicts the conclusion of the authors that the GEXP treatment was highly effective in reducing fear and avoidance beliefs in the six participants. Several considerations might explain these contradicting conclusions, which we will come back to in the Discussion section.

Illustration 3: Hypothetical data from a variety of single-case designs

In the previous illustrations, we synthesized the results of SCEs that had used the same type of single-case design. An advantage of the CICES method is that we can also synthesize the results from SCEs using different types of single-case designs. For example, suppose that a researcher wants to synthesize the results of six individual SCEs that assess the effect of a customized behavioral treatment on depression. The six SCEs use Likert scales to measure the dependent variable, but with different numbers of points on the scale. Furthermore, the studies consist of various single-case designs: two studies using an AB design with a minimum phase length of five observations, one study using an ABAB design with a minimum phase length of three observations, two studies using an alternating-treatment design (ATD) with a maximum of two consecutive administrations of the same condition, and one study using a randomized-block design (RBD). Figure 3 displays the data for the six studies.

The AB designs used in Studies 1 and 2 are among the most commonly used designs in single-case research (Shadish & Sullivan, 2011). The ABAB design used by Study 3 is an extension of the basic AB design and features two separate AB phase pairs. In both the AB design and the ABAB design, the random assignment of measurement occasions to treatment conditions pertains to the starting point of the intervention (Onghena, 2005). In the AB design there is only one moment of phase change, and thus only one random starting point for the intervention. In the ABAB design there are three moments of phase change, and thus three random starting points for the intervention need to be selected. Studies 4 and 5 use an ATD, which does not feature separate phases, but rather quick alternations of the experimental condition. Study 6 uses an RBD, which groups measurement occasions in pairs and randomizes the treatment order within each pair (Onghena, 2005). Due to the different types of single-case designs and numbers of measurement occasions, each of these hypothetical SCEs has a different set of permissible assignments. The number of permissible assignments for each of the six studies is displayed in Table 7.

Consequently, the total number of permissible combined assignments is 7*3*20*84*518*64 = 1,169,602,560. As in the previous examples, we will use a random subset of these assignments for CICES, rather than the entire set.

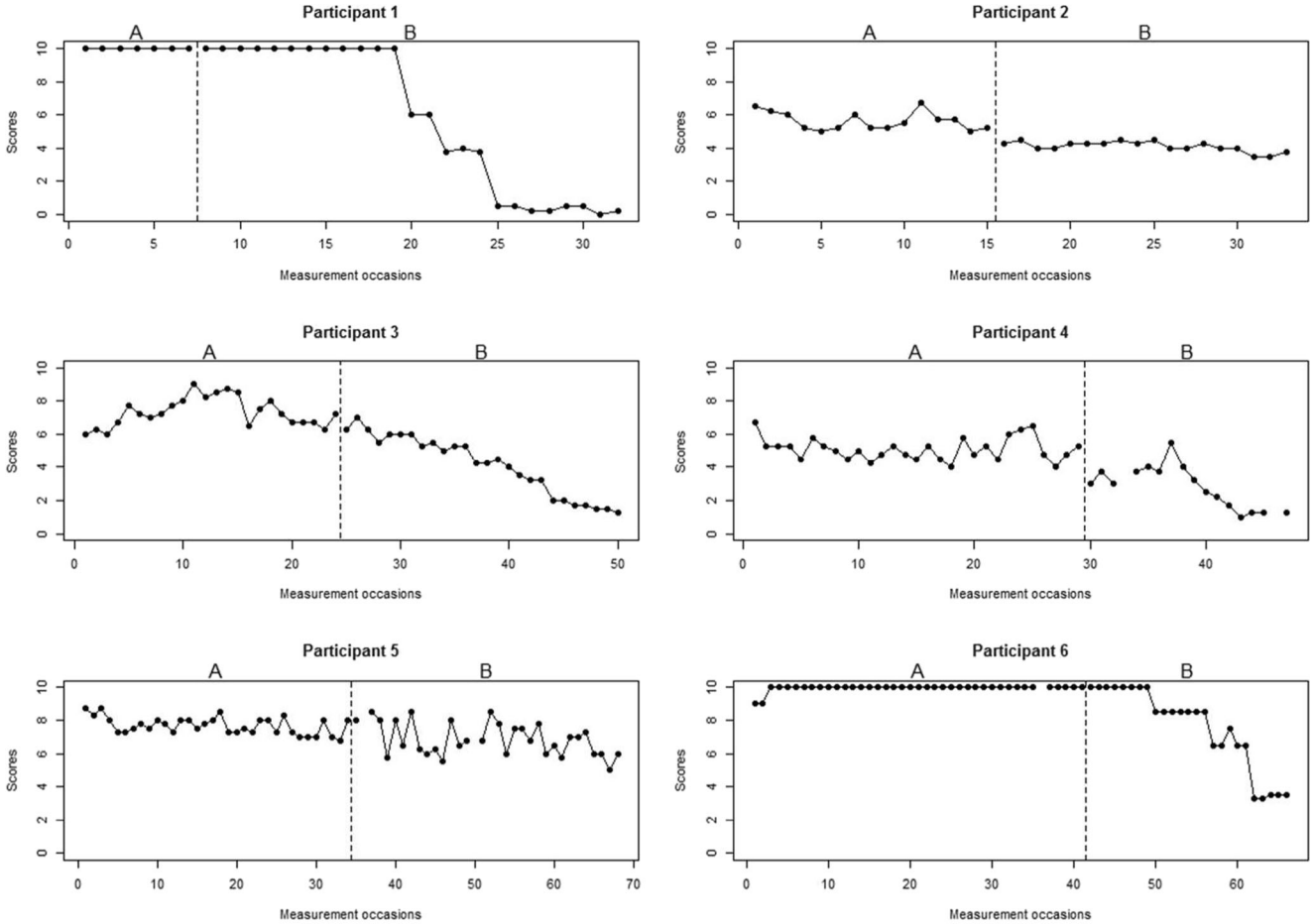

Fig. 2 Boersma et al. (2004) data for the baseline phase (A) and the treatment phase (B). Note that there are some gaps in the graphs due to missing data

Because the six SCEs use different measurement scales, and because it is unlikely that they would exhibit the same level of within-case variability, it is recommended to use a standardized ES measure for the CICES technique. Using a standardized ES measure enables us to calculate comparable ESs across the various SCEs that can be pooled into one combined ES. One possibility for the standardized ES measure is the Hedges et al. (2012) *d* statistic, which we used in the first illustration. However, this measure was specifically developed for use with AB^k phase designs, and it is thus not suited for use with single-case alternation designs. For this reason, we will use Hedges's (1981) *g* statistic with correction for small sample size bias as an alternative standardized ES measure in this example. Note that to obtain the actual test statistic in the CICES procedure, we calculate *g* separately for each individual study and then average these values. The observed value of this test statistic is 2.91. Using a precision of two decimal places, a significance level of 5%, and a Monte Carlo RT with 5,000 random assignments, we obtain a 95% confidence interval of [0.52 ; 5.31] for the average *g* across all six individual studies. The absence of the value 0 from the confidence interval indicates that the customized behavioral treatment has a significant average effect across the six individual studies at the 5% significance level. The corresponding two-sided *p* value is .0002. Although the different studies use different types of randomization schemes, the reported confidence interval is still valid, since the randomization scheme of each individual study is respected in the CICES technique. The width of the constructed nonparametric confidence interval functions as a measure of uncertainty about the observed average *g*.

As a final remark pertaining to the three illustrations, note that the average ESs are quite large and that the 95% confidence intervals are wide. The width of the intervals is due to the fact that CICES averages the effect sizes of all the individual studies into one effect size. This averaged effect size is used for statistical inference but may result in wide confidence intervals if the individual effect sizes vary considerably between studies. Furthermore, we should stress that *g* values in single-case research are not directly comparable to *g* values in a between-subjects design, because the denominator of *g* represents two different things in these contexts (Shadish et al., 2014). As a result, interpretational guidelines for the Hedges *g* in between-subjects research are not applicable to single-case research.

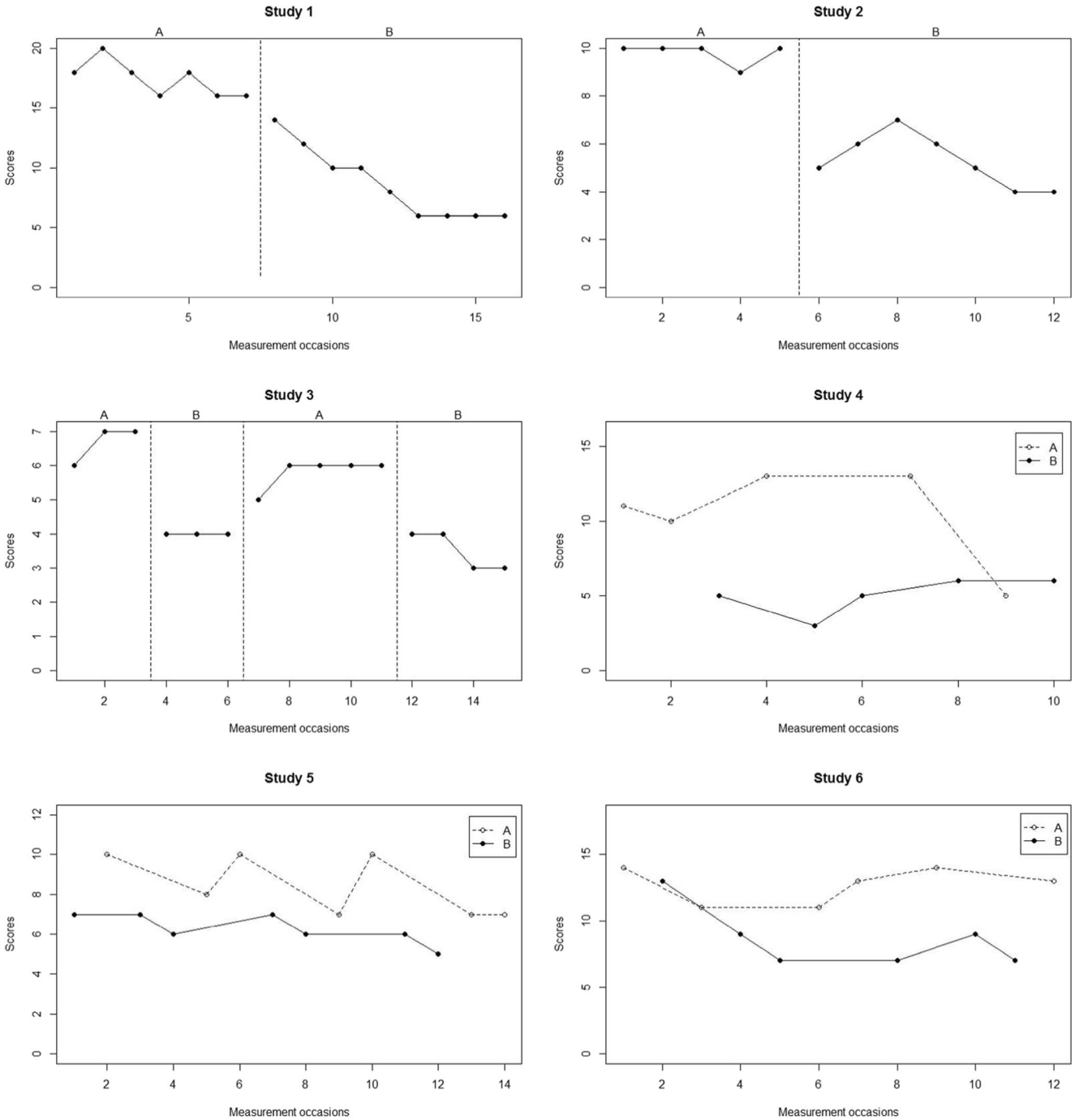

Fig. 3 Hypothetical data for six single-case experiments using various experimental designs. For each experiment, the baseline measurements are marked by the letter A, and the treatment measurements are marked by the letter B

Discussion

In this article we have presented CICES as a nonparametric meta-analytic technique for randomized SCEs and provided illustrations with empirical data as well as hypothetical data. In this section we will first recapitulate the main features of the CICES technique. Second, we will make a few additional remarks regarding the three illustrations of CICES in this article. Third, we will discuss some considerations regarding the statistical conclusion validity of the CICES technique. Fourth, we will discuss the limitations of CICES and talk about possibilities for future research.

CICES uses a random assignment model to construct a nonparametric CICES by inverting repeated randomization tests. A first characteristic of the CICES technique for single-case meta-analysis is that no distributional assumptions or an assumption of random sampling have to be made in order to make valid statistical inferences. A second

Table 7 Numbers of permissible assignments (PA) for six hypothetical single-case experiments

Study	PA
1	7
2	3
3	20
4	84
5	518
6	64

characteristic of the technique is that it can be applied to the different types of single-case ES measures that are used throughout the literature. A third characteristic of CICES as a meta-analytic technique is that it can be used to synthesize single-case experiments that use different types of randomized single-case designs (e.g., phase designs, alternation designs, etc.).

In the empirical illustration of the CICES technique for the Lambert et al. (2006) data we calculated a nonparametric confidence interval for Hedges et al.'s (2012) *d* statistic. Here we will compare the *d* confidence interval constructed by CICES to the *d* confidence interval reported by Shadish et al. (2014). The confidence interval reported by Shadish et al. is constructed by estimating the variance of *d* and using this value to construct confidence bounds for the observed *d* statistic. Hedges et al. (2012) note several distributional assumptions that have to be made in order to enable statistical inference for their *d* statistic such as the homogeneity of within-case errors and between-case variation, an AR1 autocorrelation structure for the within-case errors and normality for the between-case errors. In contrast, using the CICES technique, one does not have to make these assumptions in order to construct confidence intervals for Hedges et al.'s (2012) *d* statistic or any other ES measure for that matter. When comparing the two types of confidence intervals, one can see that the 95% confidence interval produced by the CICES technique is wider than the 95% confidence interval reported by Shadish et al. This extra width can be regarded as the cost of giving up on the distributional assumptions that the confidence interval reported by Shadish et al. required. In this sense, we argue that both types of confidence intervals for Hedges et al.'s (2012) *d* statistic can be complementary, depending on the assumptions one wishes to make about the data.

In the second empirical illustration of the CICES technique, we calculated a 95% nonparametric CICES of the Boersma et al. (2004) data. The authors' conclusion that there was a substantial effect of the GEXP treatment on fear–avoidance conflicts with the fact that a null effect was included in the 95% CICES we calculated. Several reasons might explain these contradictory conclusions. First of all, the authors did not statistically analyze the data, but only performed visual analysis. Although visual analysis is obvious and straightforward when analyzing single-case data, it is recommended to complement the visual analysis with a statistical analysis in order achieve greater certainty about the efficacy of the treatment. Second, the authors' visual analysis might have focused on individual subjects, whereas the CICES technique looks at the combined ES across all subjects. Third, we did not include the data from the posttest phase in the analysis, because this phase did not entail an experimental manipulation. However, the fear–avoidance scores still decreased considerably in the posttest phase, which might have influenced the authors' judgment with respect to the visual analysis. Fourth, the data in Fig. 2 show striking trends and delayed changes in level during the treatment phase. Such data characteristics might have given an impression of substantial impact in Boersma et al.'s visual analysis, but they are not sufficient to demonstrate a convincing causal relationship between treatment and outcome in an inferential procedure such as the CICES technique. Finally, it is also possible that the unit-treatment additivity model that was used for this analysis might not hold for these particular data, since they contain trends and delayed changes in level. As such, the results from CICES for this example should be interpreted cautiously. We will come back to this issue in the Limitations section.

In the third empirical illustration, we demonstrated that the CICES technique can be used to synthesize the raw data of SCEs that use different experimental designs. This is an important characteristic of the CICES technique, because different studies that investigate the same type of treatment sometimes use different types of single-case designs.

Throughout the illustrations in this article, we also showed that the CICES technique can be used with different single-case ES measures. This is also a handy feature of the CICES technique, because a plethora of different single-case ES measures are currently being used in the literature, and at this moment there is no widespread consensus about which ES measure is optimal for synthesizing the results of multiple SCEs (Kratochwill et al., 2010; Parker et al., 2011; Wolery et al. 2010). In addition, some proposed ESs for SCEs are not based on formal statistical distribution theory and therefore cannot be used with common parametric tools for meta-analysis (e.g., fixed or random effects meta-analysis; Shadish et al., 2014). However, all of these ES measures can be used with the CICES technique because the statistical reference distribution is derived from the randomization model, not from specific distributional assumptions.

CICES can be used with various types of test statistics, including standardized mean differences such as Hedges *g*. However it is important to note that standardized effect sizes in single-case research are not directly comparable to standardized effect sizes in between-subjects research (Shadish

et al., 2014). As such, interpretational guidelines for standardized effect sizes that were originally developed for between-subjects research (e.g., Cohen's *d*) do not apply to single-case research. Consequently the need arises to develop appropriate interpretational guidelines for standardized effect sizes in single-case research. For example, Robey and Beeson (2005) provide tentative benchmarks for standardized single-case effect sizes for a treatment that focuses on syntactic production in the aphasia literature. Endeavors such as these are highly necessary to make informed interpretations about treatment effect size magnitude in various domains of single-case research.

Limitations

CICES has a few limitations that we will now address:

First of all, and most importantly, the nonparametric confidence intervals based on the CICES method only have guaranteed validity if all individual studies included in the analysis are randomized SCEs. This is a nontrivial requirement as a considerable number of SCEs in published research are not randomized. Incorporating random assignment into an SCE increases the internal validity of the SCE substantially (e.g., Cook & Campbell, 1979; Edgington & Onghena, 2007; Heyvaert, Wendt, Van den Noortgate, & Onghena, 2015; Kratochwill & Levin, 2010; Shadish, Cook, & Campbell, 2002). Furthermore, recently published single-case reporting guidelines such as SCRIBE (Tate et al., 2016) and CENT (Vohra et al., 2015) note random assignment as an important methodological quality indicator for SCEDs. Random assignment strengthens the SCE's internal validity because it yields statistical control over confounding variables such as history and maturation (Levin & Wampold, 1999; Onghena, 2005).

Although random assignment is obviously very important for the methodological quality of a single-case design, it is not the case that nonrandomized studies are completely useless for inference. First, when random assignment is absent an overall treatment effect can still be detected by CICES (along with a corresponding confidence interval) but then one cannot unambiguously attribute this overall treatment effect to the experimental manipulations in the SCEs. In other words, for nonrandomized SCEs the results of CICES can only be interpreted descriptively and not inferentially. Second, when RTs or CICES are used for nonrandomized designs, nominal Type I error rates are not guaranteed (Ferron, Foster-Johnson, & Kromrey, 2003). See Winch and Campbell (1969) for a more thorough discussion on the issue of using randomization tests for nonrandomized designs. Given these considerations we recommend caution in the interpretation of the results when using CICES for nonrandomized designs.

Some authors are opposed to the practice of randomizing single-case designs. For example, one of the arguments presented against the use of randomization tests for analyzing SCEs is that response-guided experimentation becomes impossible (e.g., Kazdin, 1980). The argument is that in single-case research decisions to implement, withdraw, or alter treatments are often based on the observed data patterns during the course of the experiment. However, randomization tests require determining in advance and in a random fashion when the treatment will be implemented, thus making response-guided adjustments to the experiment impossible. Edgington (1980) responded to this criticism by proposing an RT in which only part of the measurement occasions of the SCE are randomized and thus gives control to the researcher over the nonrandomized part. As another point of criticism, Kazdin (1980) argues that the randomization requirements of the RT are often at odds with the practical feasibility of an SCE in a clinical context. For example, the administration of a treatment during an SCE might require administrative support and special monitoring procedures from several staff members. If the times at which the treatment is administered is determined randomly, it is likely that logistic problems will occur with respect to the availability of the required staff and equipment for the proper administration of the treatment at that time. Given these considerations we would recommend to randomize SCEs whenever the practical consequences of randomization do not form an obstacle for conducting the SCE.

In practice it is likely that single-case meta-analysts will have a mix of randomized and nonrandomized studies. Strictly speaking, when CICES is used for a set of studies that include at least one nonrandomized study the validity of the overall inference is potentially compromised. One solution for this problem would be to do separate meta-analyses for randomized SCEs and nonrandomized SCEs. Furthermore one could perform a sensitivity analysis by comparing the results of both groups of SCEs. If the results for both groups of SCEs are similar, it is more plausible that there were no major confounding variables at play in the nonrandomized studies.

The validity of the nonparametric confidence interval produced by CICES is also based on the assumption that the unit-treatment additivity model (see above) is an accurate conceptualization of the treatment effect. Hence, an important question then is whether the unit-treatment additivity model provides an accurate description of a treatment effect in single-case data. Indeed the assumption of a constant additive treatment effect might not be tenable in every type of research situation. For example, research has shown that single-case data can contain time-related effects such as serial correlation (e.g., Matyas & Greenwood, 1997; Shadish & Sullivan, 2011) and trends (e.g., Beretvas & Chung, 2008; Manolov & Solanas, 2009; Solomon, 2014). In such a situation, it is possible that the onset of the treatment interacts with these time-related effects. For example, the onset of the treatment might instigate a trend change relative to the baseline phase (Van den Noortgate & Onghena, 2003) or induce a change in score variability (Ferron, Moeyaert, Van den Noortgate, &

Beretvas, 2014). These unit-treatment interactions are not accounted for in the unit-treatment additivity model and may thus confound treatment effect estimates when such interactions are present.

One way to account for time-related effects in the data would be to incorporate a time parameter in the unit-treatment additivity model along with a parameter for the expected time-related effect. For example one could formulate an effect size model that accounts for deterministic trends in the null scores: $Y_i = (X_i + t\beta) + D_i\Delta$ with t being a time variable and β being a constant trend effect. The null hypothesis for this model is $H_0 : Y_i = (X_i + t\beta)$. This model could be used to evaluate a treatment for a patient that is expected to exhibit spontaneous recovery in the outcome variable. Alternatively one could devise a model in which the treatment causes a mean level shift as well as a deterministic trend: $Y_i = X_i + D_i(\Delta + t_B\beta)$ with t_B being the time variable for the treatment phase. In this case the null hypothesis is $H_0 : Y_i = X_i$. The inclusion of a time variable t into the model could also account for delayed treatment effects. Instead of additive models one could also consider multiplicative models that assume a non-linear relation between the null scores and the observed scores. For example one could formulate a model in which the magnitude of the treatment effect for experimental unit i is inversely related to its null score: $Y_i = X_i + D_i \frac{1}{X_i}\Delta$. Generally speaking, the difference between a set of null scores and a set of observed scores can be modeled using any type of function f as long as the equation $Y_i = f(X_i)$ holds. In other words the effect size function should be a bijective function between the null scores and the observed scores.

The previously mentioned models are all deterministic: Given a function $f()$, there is a perfect correspondence between the null scores and the observed scores. However the effect model can also incorporate random effects. One example is the extended unit-treatment additivity model (Cox & Reid, 2000): $Y_i = X_i + D_i\Delta + \varepsilon_i$. In this model the ε_i are independent and identically distributed random variables with a mean of zero and a variance of σ. Of course in this model the distributional characteristics of the ε_i must be specified, which invokes a distributional assumption in the CICES procedure. Random effects can be used to allow for variations in treatment effect size across j included studies in CICES: $Y_{ij} = X_{ij} + D_{ij}(\Delta + \varepsilon_j) + \varepsilon_i$.

These examples illustrate that CICES actually has a tremendous flexibility with regard to the effect size functions that can be used for the statistical inference. However it should be noted that the effect size function that is chosen prior to conducting the experiment must be plausible and well interpretable. Note also that misspecification of the effect size function can severely diminish the power of the underlying RT. With regard to the use of the unit-treatment additivity model in the current implementation of CICES we want to emphasize that this model is also implicitly used in standard parametric tests such as t tests and F tests, which are basically significance tests for detecting mean level shifts between experimental groups.

That being said we can state that the use of CICES with the unit-treatment additivity model is most appropriate for situations in which the data is not expected to contain large trends and/or changes in variability or other types of effects that might indicate the presence of unit-treatment interactions. When effects such as trends and changes in variability are expected an alternative effect size model can be used that takes these effects into account. In this sense the a priori choice of an appropriate effect size model for CICES is as important as other choices that have to be made for any valid statistical analysis, such as the choice of the research design, the number of observations, and the test statistic. It goes without saying that optimal choices for these design parameters depend on the research question, the predicted effects, and the statistical power of the test.

Future research

In light of the discussion about the tenability of the unit-treatment additivity model for single-case meta-analysis, one avenue for future research could be to investigate the influence of effect size model misspecification on the Type I error and power of CICES. This could be done by means of a simulation study in which data for a group of SCEs is first generated using a specific effect size model (e.g., unit-treatment additivity with a linear trend component) and then evaluated using CICES but with a different effect size model (e.g., unit-treatment additivity model).

In addition, future research could expand the CICES technique to other single-case ES measures. For example, CICES might be expanded to include ES measures based on data overlap (see Parker et al., 2011, for an overview) or ES measures that are sensitive to effects other than mean level shifts (e.g., trends, changes in variability, etc.). The CICES technique is very flexible in this regard but one consideration to keep in mind is that the underlying effect model of CICES must be compatible with the ES measure that is being used in the RT. More specifically, the type of treatment effect (e.g., mean level difference effect, trend change, or variability change) that the selected ES is designed to measure must also be used to construct the "observed scores" from the "null scores." This means that effect models other than the unit-treatment additivity model are probably more appropriate to model these alternative types of effects.

Another avenue for further research would be the use of the CICES technique to synthesize results from randomized between-subjects designs with results from randomized single-case designs. Most between-subjects designs incorporate the random assignment of experimental units to treatment conditions, just like randomized SCEs. The difference

between between-subjects designs and SCEs in this respect is that the experimental units in between-subjects designs pertain to individual persons whereas the experimental units in SCEs pertain to repeated measurements within a single person. However, this would bring along a series of conceptual considerations. One challenge in this approach would be how to adequately weigh the raw data in the calculation of the combined ES because between-subjects designs contain only one data point per participant whereas all data points in a single SCE originate from a single participant. Another conceptual issue would be if the resulting combined ES estimate and confidence interval can be interpreted meaningfully when it pertains to a mix of single-case designs and between-subjects designs.

Software availability

We have developed an R function for the CICES technique that is freely available from https://ppw.kuleuven.be/home/english/research/mesrg/appletsandsoftware. A page on this website is dedicated to CICES, containing a .zip file with the R code and the datasets we used in this article, as well as user instructions. We strongly recommend to use the R function in the R studio graphical user interface for the R programming language. Although the R function automatically performs all the CICES calculations, some basic knowledge of R is required to properly set up and use the code. The function in its current form employs the unit-treatment additivity model as described in this article. Although other effect size functions can also be used some additional adjustments to the R code would be necessary to implement them.

Conclusion

In this article we have introduced the CICES technique to calculate nonparametric confidence intervals for the combined ES of multiple randomized SCEs. CICES is based on the inversion of an RT and offers tremendous flexibility with respect to the types of ES measures and single-case experimental designs it can handle. Furthermore, the technique requires no distributional assumptions and no assumption of random sampling. Importantly, the validity of CICES is dependent on the assumption that the included SCEs have randomized designs. Furthermore, the suitability of CICES for specific data patterns that are present in the included datasets is dependent on the specific effect size model that is used. In the current implementation of CICES we used the unit-treatment additivity model, which assumes a constant additive treatment effect between the null scores and the observed scores across all included studies. Consequently the unit-treatment additivity model is best used for data that do not contain trends or changes in variability throughout the SCE and for groups of SCEs that are not characterized by large treatment effect size heterogeneity. Future research should focus on the use of more complex effect size models in CICES for analyzing data sets with various unit-treatment interaction effects. Because CICES is a novel technique, future research is needed to further validate this technique and investigate its practical feasibility for single-case researchers as well as its applicability to various real-world data-analytical situations. We hope that the CICES technique can be of value to single-case researchers for meta-analyzing single-case experiments.

Author note This research was funded by the Research Foundation–Flanders (FWO), Belgium (Grant ID: G.0593.14)

References

American Psychological Association. (1994). Publication manual of the American Psychological Association (4th ed.). Washington, DC: Author.

American Psychological Association. (2001). Publication manual of the American Psychological Association (5th ed.). Washington, DC: Author.

American Psychological Association. (2010). Publication manual of the American Psychological Association (6th ed.). Washington, DC: Author.

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

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.

Boersma, K., Linton, S., Overmeer, T., Jansson, M., Vlaeyen, J., & De Jong, J. (2004). Lowering fear-avoidance and enhancing function through exposure in vivo: A multiple baseline study across six patients with back pain. *Pain*, *108*, 8–16.

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

Burns, M. K. (2012). Meta-analysis of single-case design research: Introduction to the special issue. *Journal of Behavioral Education*, *21*, 175–184.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*, 1304–1312. doi:https://doi.org/10.1037/0003-066X.45.12.1304

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003. doi:https://doi.org/10.1037/0003-066X.49.12.997

Cook, T. D., & Campbell, D. T. (1979). Quasi-experimentation: Design and analysis issues for field settings. Chicago, IL: Rand McNally.

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

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7–29. doi:https://doi.org/10.1177/0956797613504966

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

Dwass, M. (1957). Modified randomization tests for nonparametric hypotheses. *Annals of Mathematical Statistics*, *28*, 181–187.

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

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

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

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

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

Ferron, J. M., Bell, B. A., Hess, M. F., Rendina-Gobioff, G., & Hibbard, S. T. (2009). Making treatment effect inferences from multiple-baseline data: The utility of multilevel modeling approaches. *Behavior Research Methods*, *41*, 372–384.

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

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

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

Ferron, J. M., Moeyaert, M., Van den Noortgate, W., & Beretvas, S. N. (2014). Estimating casual effects from multiple-baseline studies: Implications for design and analysis. *Psychological Methods*, *19*, 493–510.

Garthwaite, P. (2005). Confidence intervals: Nonparametric. In B. S. Everitt & D. C. Howell (Eds.), Encyclopedia of statistics in behavioral science (pp. 375–381). Chichester, UK: Wiley.

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

Hedges, L. G., Pustejovsky, J. E., & Shadish, W. R. (2012). A standardized mean difference effect size for single-case designs. *Research Synthesis Methods*, *3*, 224–239.

Hedges, L. G., 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.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*, 107–128.

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

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

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

Hinkelmann, K., & Kempthorne, O. (2005). Design and analysis of experiments, *Vol. 2: Advanced experimental design*. Hoboken, NJ: Wiley.

Hinkelmann, K., & Kempthorne, O. (2008). Design and analysis of experiments, Vol. 1: Introduction to experimental design (2nd ed.). Hoboken, NJ: Wiley.

Hinkelmann, K., & Kempthorne, O. (2012). *Design and analysis of experiments, Vol. 3: Special designs and applications*. Hoboken, NJ: Wiley.

Hope, A. C. A. (1968). A simplified Monte Carlo test procedure. *Journal of the Royal Statistical Society: Series B*, *30*, 582–598.

Horner, R. D., & Baer, D. M. (1978). Multiple probe technique: A variation on the multiple baseline. *Journal of Applied Behavior Analysis*, *11*, 189–196.

Kazdin, A. E. (1980). Obstacles in using randomization tests in single-case experimentation. *Journal of Educational Statistics*, *5*, 253–260.

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

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement*, *56*, 746–759.

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

Kratochwill, T. R., & Levin, J. R. (2010). Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. *Psychological Methods*, *15*, 124–144. doi:https://doi.org/10.1037/a0017736

Kratochwill, T. R., & Levin, J. R. (Eds.). (2014). Single-case intervention research: Methodological and statistical advances. Washington, DC: American Psychological Association.

Lambert, M. C., Cartledge, G., Heward, W. L., & Lo, Y. (2006). Effects of response cards on disruptive behavior and academic responding during math lessons by fourth-grade urban students. *Journal of Positive Behavior Interventions*, *8*, 88–99.

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

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.

Levin, J. R., Marascuilo, L. A., & Hubert, L. J. (1978). *N* = Nonparametric randomization tests. In T. R. Kratochwill (Ed.), Single-subject research: Strategies for evaluating change (pp. 167–196). New York, NY: Academic Press.

Levin, J. R., & Wampold, B. E. (1999). Generalized single-case randomization tests: Flexible analyses for a variety of situations. *School Psychology Quarterly*, *14*, 59–93.

Maas, C. J. M., & Hox, J. J. (2004). Robustness issues in multilevel regression analysis. *Statistica Neerlandica*, *58*, 127–137.

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

Manolov, R., & Solanas, A. (2009). Percentage of nonoverlapping corrected data. *Behavior Research Methods*, *41*, 1262–1271. doi: https://doi.org/10.3758/BRM.41.4.1262

Matyas, T. A., & Greenwood, K. M. (1997). Serial dependency in single-case time series. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), Design and analysis of single-case research (pp. 215–243). Mahwah, NJ: Erlbaum.

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.

Moeyaert, M., Ugille, M., Ferron, J. M., Beretvas, S. N., & Van den Noortgate, W. (2014). Three-level analysis of single-case experimental data: Empirical validation. *Journal of Experimental Education*, *82*, 1–21.

Moore, D. S., McCabe, G. P., & Craig, B. A. (2014). Introduction to the practice of statistics (8th ed.). New York, NY: W. H. Freeman.

Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society A*, *767*, 333–380.

Nugent, W. (1996). Integrating single-case and group comparison designs for evaluation research. *Journal of Applied Behavioral Science*, *32*, 209–226.

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

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

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

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

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

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

Parker, R. I., Vannest, K. J., & Davis, J. L. (2011). Effect size in single-case research: A review of nine nonoverlap techniques. *Behavior Modification*, *35*, 303–322.

Rindskopf, D. M. (2014). Nonlinear Bayesian analysis for single case designs. *Journal of School Psychology*, *52*, 71–81.

Robey, R. R., & Beeson, P. M. (2005). Aphasia treatment: Examining the evidence. Presentation at the American Speech-Language-Hearing Association Annual Convention. San Diego, CA.

Scruggs, T. E., & Mastropieri, M. A. (2013). PND at 25: Past, present, and future trends in summarizing single-subject research. *Remedial and Special Education*, *34*, 9–19.

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.

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

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi experimental designs for generalized causal inference. New York, NY: Houghton Mifflin.

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.

Shadish, W. R., & Rindskopf, D. M. (2007). Methods for evidence-based practice: Quantitative synthesis of single-subject designs. *New Directions for Evaluation*, *113*, 95–109.

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

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.

Tate, R. L., Perdices, M., Rosenkoetter, U., Shadish, W., Vohra, S., Barlow, D. H., . . . Wilson, B. (2016). The Single-Case Reporting guideline In BEhavioural interventions (SCRIBE): 2016 statement. *Aphasiology*, *30*, 862–876.

Tritchler, D. (1984). On inverting permutation tests. *Journal of the American Statistical Association*, *385*, 200–207.

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

Vlaeyen, J. W., de Jong, J., Geilen, M., Heuts, P. H. T., & van Breukelen, G. (2001). Graded exposure in vivo in the treatment of pain-related fear: a replicated single-case experimental design in four patients with chronic low back pain. *Behaviour Research and Therapy*, *39*, 151–166.

Vohra, S., Shamseer, L., Sampson, M., Bukutu, C., Schmid, C. H., Tate, R., . . . the CENT Group. (2015). CONSORT extension for reporting *N*-of-1 trials (CENT): 2015 statement. *British Medical Journal*, *350*, h1738.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on *p*-values: Context, process, and purpose. *American Statistician*, *70*, 129–133.

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

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

Winch, R. F., & Campbell, D. T. (1969). Proof? No. Evidence? Yes. The significance of tests of significance. *American Sociologist*, *4*, 140–143.

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