

Confidence intervals for single-case effect size measures based on randomization test inversion

Bart Michiels¹ · Mieke Heyvaert¹ · Ann Meulders¹ · Patrick Onghena¹

Published online: 29 February 2016
© Psychonomic Society, Inc. 2016

Abstract In the current paper, we present a method to construct nonparametric confidence intervals (CIs) for single-case effect size measures in the context of various single-case designs. We use the relationship between a two-sided statistical hypothesis test at significance level α and a 100 $(1 - \alpha)$ % two-sided CI to construct CIs for any effect size measure θ that contain all point null hypothesis θ values that cannot be rejected by the hypothesis test at significance level α . This method of hypothesis test inversion (HTI) can be employed using a randomization test as the statistical hypothesis test in order to construct a nonparametric CI for θ . We will refer to this procedure as randomization test inversion (RTI). We illustrate RTI in a situation in which θ is the unstandardized and the standardized difference in means between two treatments in a completely randomized single-case design. Additionally, we demonstrate how RTI can be extended to other types of single-case designs. Finally, we discuss a few challenges for RTI as well as possibilities when using the method with other effect size measures, such as rank-based nonoverlap indices. Supplementary to this paper, we provide easy-to-use R code, which allows the user to construct nonparametric CIs according to the proposed method.

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

Introduction

Single-case experiments (SCEs) can be used to assess the efficacy of an intervention or treatment for a single person. In such experiments, repeated measurements are taken for a single person on a dependent variable of interest and the treatment can be considered as one of the levels of the independent variable (Barlow, Nock, & Hersen, 2009; Kazdin, 2011; Onghena, 2005). Over the years, the use of SCEs has risen steadily in disciplines such as school psychology, clinical psychology, and medical science (Hammond & Gast, 2010; Shadish & Sullivan, 2011; Swaminathan & Rogers, 2007).

SCE data are traditionally analyzed visually and this remains the primary analysis method to date (e.g., Bulté & Onghena, 2012; Gast & Ledford, 2014; Kazdin, 2011; Kratochwill et al., 2010). The main advantage of visual analysis is that aspects of SCE data such as level, trend, variability, immediacy of the effect, and overlap can be assessed in a flexible way (Horer et al., 2005; Kratochwill, Levin, Horner, & Swoboda, 2014; Lane & Gast, 2014). However, visual analysis has been criticized for its lack of established formal decision guidelines which leaves the method vulnerable to subjectivity and inconsistency between researchers (e.g., Deprospero & Cohen, 1979; Fisch, 1998; Gibson & Ottenbacher, 1988; Harrington & Velicer, 2015; Ximenes, Manolov, Solanas, & Quera, 2009).

Over the years, several types of statistical analysis methods have been proposed for SCE data with the intent to obtain more objective measures of effect size (ES) that can be used to complement the results of visual analysis. These include regression-

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

¹ Faculty of Psychology and Educational Sciences, KU Leuven - University of Leuven, Leuven, Belgium

based measures (e.g., Allison & Gorman, 1993; Center, Skiba, & Casey, 1985–1986; Solanas, Manolov, & Onghena, 2010; Van den Noortgate & Onghena, 2003; White, Rusch, Kazdin, & Hartmann, 1989), standardized mean difference measures (e.g., Busk & Serlin, 1992; Hedges, Pustejovsky, & Shadish, 2012) and measures based on data nonoverlap between phases (e.g., Parker, Hagan-Burke, & Vannest, 2007; Parker & Vannest, 2009; Parker, Vannest, & Brown, 2009; Parker, Vannest, Davis, & Sauber, 2011). These advances in developing useful measures of ES for SCE data were also propagated by the increasing demand from scholarly journals, academic organizations, and policy makers to include effect sizes (ESs) in the publication of scientific results. For example, the fourth edition of the publication manual of the American Psychological Association (1994) included a new guideline to report ESs as a standard practice in psychological research. Similarly, Wilkinson and the Task Force on Statistical Inference (1999) advocated the use of ESs in addition to the reporting of p -values in the behavioral sciences. The call for more frequent use of ESs in SCEs also stems from the increasingly popular evidence-based practice movements in school psychology (Kratochwill & Stoiber, 2000) and clinical psychology (Chambless & Ollendick, 2001).

Apart from the quantification of ES, another aspect of evaluating SCE data is statistical inference – performing statistical hypotheses tests and constructing confidence intervals (CIs) for ESs. However, the validity of statistical inferences based on parametric procedures is frequently doubtful because the assumptions underlying these procedures (e.g., random sampling or specific distributional assumptions) are implausible in many areas of behavioral research, and for single-case research in particular (e.g., Adams & Anthony, 1996; Dugard, 2014; Edgington & Onghena, 2007; Ferron & Levin, 2014; Levin, Ferron, & Gafurov, 2014; Micceri, 1989). Therefore, the randomization test (RT) has been proposed as a nonparametric alternative to test statistical hypotheses in randomized SCEs (e.g., Bulté & Onghena, 2008; Edgington, 1967; Onghena & Edgington, 1994, 2005; Heyvaert & Onghena, 2014; Levin, Ferron & Kratochwill, 2012; Onghena, 1992). The RT makes no assumption of random sampling and no distributional assumptions. Instead, the test uses the random assignment of measurement occasions to treatment conditions as the basis of its validity (Edgington & Onghena, 2007).

The current paper proposes a method to construct nonparametric CIs for mean difference type ESs in SCEs using the RT rationale. This method is developed by combining, on the one hand, the principle of *hypothesis test inversion* (HTI) (e.g., Garthwaite, 2005; Trichler, 1984) which exploits the equivalence between a hypothesis test and a CI and on the other hand, the use of RTs. Consequently, we have named this new method *randomization test inversion* (RTI). The method offers flexibility with regard to the choice of an ES measure and can be used with any randomized experimental design. The focus of this paper will be on the application of RTI to randomized single-case designs.

We hope that RTI can be of value to the applied single-case researcher for gauging the uncertainty of the size of a treatment effect. However, note that we do not propose to replace visual analysis with the sole use of nonparametric CIs for ESs. We concur with the general consensus in the field of single-case research that visual and statistical analysis are complementary and in most cases should be used together to corroborate the conclusions and to increase the acceptability by the wider scientific community (e.g., Bulté & Onghena, 2012; Busk & Marascuilo, 1992; Harrington & Velicer, 2015; Kratochwill et al., 2014; Tate et al., 2013).

In the following sections, we will elaborate on the two main components of the RTI method: (1) the principle of HTI and (2) the use of RTs, and how they fit together in order to form the proposed RTI method. We will use a step-by-step approach in which we gradually build up to the application of the proposed RTI method for single-case designs. We will start with an explanation of the HTI principle in the context of a between-subject design and illustrate how we can use this principle to invert a parametric test (a two-sample t -test in the example) in order to arrive at the same parametric CI that can be analytically derived for that parametric test. Next, we will elaborate on how the HTI principle can be combined with an RT (forming the RTI method) in order to produce nonparametric CIs for ES measures in between-subject designs. Subsequently, we will show and illustrate how the RTI method can be extended to accommodate various types of single-case designs such as completely randomized single-case designs (CRD), alternating treatments designs (ATD), randomized block designs (RBD), as well as phase designs (so-called AB, ABA, or ABAB designs). The ESs we will use in our illustrations for single-case designs consist of an unstandardized and standardized mean difference for CRDs, ATDs, and RBDs and an immediate treatment effect index for phase designs. Finally, we will discuss a few challenges for RTI as well as possibilities when the method is implemented with other ESs, such as rank-based nonoverlap indices.

Hypothesis test inversion

There exists a necessary equivalence between a two-sided hypothesis test and a two-sided CI. More specifically, a 100 $(1 - \alpha) %$ two-sided CI for a parameter θ consists of all point null values of θ that cannot be rejected by a two-sided hypothesis test at significance level α (Neyman, 1937). From this equivalence, one can construct CIs for θ by repeatedly performing hypothesis tests on a range of hypothesized values of θ and including the nonrejected values as part of the CI. We will refer to this procedure as the *method of hypothesis test inversion* (HTI).

HTI goes as follows: Let θ denote the parameter for which we want to construct a two-sided CI. We can then choose a

range of point null values (θ_0) for which we test the null hypothesis $\theta = \theta_0$ against the two-sided alternative hypothesis that $\theta \neq \theta_0$. If α is the significance level of the two-sided hypothesis test, we can construct a 100 $(1 - \alpha)$ % CI for the parameter θ that contains all θ_0 values for which the null hypothesis is not rejected (Garthwaite, 2005; Tritchler, 1984).

To introduce HTI, we will start with the situation in which we want to construct a parametric CI for a mean difference based on a classical t -procedure in the context of a between-subject experimental design. We will subsequently show that inverting the t -test through HTI results in exactly the same parametric CI as the one that can be computed analytically.

Confidence intervals (CIs) for completely randomized between-subject designs using hypothesis test inversion

Suppose a therapist wants to test a new experimental treatment to improve quality of life in patients with chronic pain using a randomized experiment. Assume that this experiment consists of a between-subject design with 16 patients, eight of whom are randomly assigned to the control condition and eight of whom are randomly assigned to the experimental condition. This type of design is also known as a completely randomized design (CRD). Let us assume the therapist is interested in the population parameter θ that represents the difference in means between both conditions. Let the variables A and B contain the scores (measured on a rating scale from 1 to 10) of the control condition and the experimental condition respectively. Now consider the hypothetical data as shown in Table 1.

The mean quality-of-life scores are 4.625 and 8 for A and B , respectively, and the observed difference in means $\theta_{obs} = \bar{B} - \bar{A} = 3.375$. If we are willing to assume that the scores represent independent random samples from two normally distributed populations with equal variances, we can test the null hypothesis that $\theta_0 = 0$ against the alternative that $\theta \neq \theta_0$ with a two-tailed parametric t -test. This test gives a t -value of 4.6208, with 14 degrees of

freedom, and a resulting two-sided p -value of .0004, indicating a statistically significant difference in means of A and B for a 5 % significance level.

We can construct a parametric CI for the difference in means by using analytically derived formulas that can be found in any introductory statistics textbook (e.g., Mann, 2006). A t -procedure for these data results in a 95 % CI of [1.81–4.94]. With an observed mean difference of 3.375, the t -procedure shows that we can be 95 % confident that the true mean difference lies between 1.81 and 4.94. This means that if the therapist would repeat the experiment a large number of times and subsequently analyze the resulting samples, the true parameter value will be contained in the CI constructed by the t -procedure in 95 % of the times (Moore, McCabe, & Craig, 2014).

The same parametric 95 % CI can also be constructed using HTI. To obtain the boundaries of the CI, we repeatedly perform t -tests for a range of θ_0 values, starting from the observed $\theta_{obs} = 3.375$ and going upward. For each θ_0 value, the observed test statistic is compared to a t -distribution with 14 degrees of freedom at the 5 % level. The upper boundary of the CI is found as the largest θ_0 that is not rejected by the t -test; the lower boundary is just at equal distance at the other side of θ_{obs} . Formally, calculate the difference θ_{diff} between this largest θ_0 and θ_{obs} , that is $\theta_{diff} = \text{largest } \theta_0 - \theta_{obs}$. The CI is defined as [Lower boundary; Upper boundary] with

$$\begin{aligned} \text{Lower boundary} &= \theta_{obs} - \theta_{diff} \\ \text{Upper boundary} &= \theta_{obs} + \theta_{diff} \end{aligned}$$

Table 2 illustrates the process of inverting the t -test for the hypothetical data of the chronic pain patients.

Obviously, if we test for $\theta_0 = \theta_{obs} = 3.375$, then the numerator of the t -statistic becomes 0, which is in the middle of the sampling distribution, and so the null hypothesis is not rejected. Therefore, the value of 3.375 is included in the CI. By gradually increasing θ_0 , we can search for the boundaries of the CI (see Table 2). For a 95 % CI the boundaries are constructed using the largest θ_0 value for which a two-tailed t -test provides a p -value that is larger than .05. For the example data, this results in the following CI:

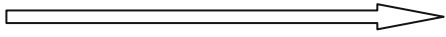
$$\begin{aligned} \text{Lower boundary} &: 3.375 + 1.565 = 1.81 \\ \text{Upper boundary} &: 3.375 - 1.565 = 4.94 \end{aligned}$$

The obtained CI is exactly the same as the analytically derived CI according to the t -procedure, as it should be. The boundaries of the CI can be reached at any desired level of precision, conditional on the granularity of the θ_0 range (Garthwaite, 2005).

Table 1 Hypothetical quality-of-life scores of 16 patients with chronic pain in a completely randomized design, comparing a control group (A) to a treatment group (B)

	A	B
	7	9
	5	8
	7	7
	4	9
	5	8
	3	6
	4	9
	2	8

Table 2 Using hypothesis test inversion to construct a parametric 95 % confidence interval (CI) for the data in Table 1



θ_{diff}	0	0.005	...	1.565	1.570
θ_0	3.375	3.380	...	4.940	4.945
p	1	.99460502	.0496
In CI	YES	YES	YES	YES	NO

Nonparametric CIs for completely randomized between-subject designs

Nonparametric CI for an unstandardized mean difference

Parametric tests such as the t -test from the previous example depend on classical assumptions such as random sampling and specific population distributions. In case of one or more of these assumptions being possibly violated, the therapist can analyze the data with a randomization test (RT), which assumes a random assignment model. RTs obtain their validity by randomly assigning experimental units to experimental conditions. By randomizing the condition labels referring to the levels of the independent variable, the null hypothesis that there is no differential effect of the levels of the independent variable on the dependent variable can be tested (Edgington & Onghena, 2007). We will refer to a specific randomization of the condition labels as an *assignment*.

Heyvaert and Onghena (2014) describe the necessary steps for performing RTs on experimental data. The first step must be made prior to executing the experiment and consists of listing all *permissible assignments* for the chosen experimental design. A permissible assignment is an assignment that adheres to the restrictions imposed by the chosen randomization scheme. An example of a randomization restriction for a completely randomized between-subject design could be that the same number of subjects is assigned to each of the conditions. The set of permissible assignments is defined by the choice of a specific experimental design (Onghena & Edgington, 2005). After listing all permissible assignments, one of them is randomly selected as the assignment for the actual experiment. Next, one must choose a test statistic that is adequate to answer the research question. For example, when one is interested in the overall effect of the levels of the independent variable on the dependent variable, one can use a mean difference statistic to quantify this effect. Note that RTs can be one-sided or two-sided depending on whether the chosen test statistic is sensitive to the direction of the alternative hypothesis. For a two-sided RT, a nondirectional test statistic (e.g., an absolute mean difference) has to be used.

The next step involves constructing the randomization distribution by computing the value of the test statistic for all

permissible assignments. The randomization distribution functions as a reference distribution that can be used to determine the statistical significance of the observed test statistic: The two-sided p -value of an RT is calculated as the proportion of (absolute value) test statistics in the randomization distribution that are at least as extreme as the observed test statistic. Depending on the chosen significance level, the therapist then either rejects or accepts the null hypothesis based on the p -value. Because the observed value of the test statistic also stems from one of the permissible assignments, the smallest achievable p -value with an RT is equal to the inverse of the number of permissible assignments (Onghena & May, 1995). There are several software packages designed to perform RTs. For example, Huo and Onghena (2012) developed a Windows-based program for performing RTs. Software for RTs is also available from the “coin” R package (Hothorn, Hornik, van de Wiel, & Zeileis, 2008).

Performing the two-sided RT for the chronic pain example yields a p -value of .0014 indicating a statistically significant difference between the means of A and B, so we reject the null hypothesis that there is no differential treatment effect on the quality-of-life scores. Notice that this p -value is only slightly larger than the p -value of the parametric t -test, but at the same time that we are operating within a nonparametric framework, without making any distributional assumptions and without making an assumption of random sampling (Edgington & Onghena, 2007).

It is also possible to perform an RT for effects other than a null effect. This can be done by assuming that the experimental condition has a constant additive effect (denoted by Δ) on the scores of the outcome variable. This model is called the “unit-treatment additivity model” and it is the model that is most popular and well studied within nonparametric statistics (e.g., Cox & Reid, 2000; Hinkelmann, & Kempthorne, 2008; Lehman, 1959; Welch & Gutierrez, 1988). The “unit-treatment additivity model” expresses the observed scores as:

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

where X_i^B is the observed score of experimental unit i if i is assigned to the experimental condition B, X_i^A is the hypothetical score of i if i instead was assigned to the control condition

A (i.e., the *null* score or the score that would be observed if the null hypothesis of no treatment effect is true), and Δ is the constant additive effect of the treatment. Note that X_i^B and X_i^A can never be observed simultaneously and that Δ has no index, that is, Δ is assumed to be constant for every experimental unit. When one performs the RT for a null effect of the experimental manipulation, Δ equals zero. In this case:

$$X_i^B = X_i^A$$

implying that under the null hypothesis that the experimental treatment has no effect, the observed score for experimental unit i is independent from the condition to which it is assigned. One can test for non-null effects by specifying a non-zero value for Δ .

Returning to the HTI method, we can use the model of unit-treatment additivity to compute a CI for a mean difference by inverting the RT. This boils down to performing the RT for a range of Δ -values and retaining the non-rejected Δ -values to construct the nonparametric CI. We will further refer to this method as *randomization test inversion* (RTI). In this context, Δ is actually the equivalent of the point null parameter value (θ_0) that we used in the introductory section on HTI. Note that this RTI method assumes that the unit-treatment additivity model is adequate to capture the treatment effect if the null hypothesis is false. Other models can be conceived of, but this would require the development of an alternative procedure. Furthermore, note that the choice for the unit-treatment additivity model implies that certain measures of ES are more obvious candidates to be used in RTI than others. In this respect we should emphasize that the unit-treatment additivity model that we adopt here is most suitable for measures of ES that indicate mean level separation. We will come back to this issue in the [Discussion](#) section where we will also critically reflect on the adequacy of the unit-treatment additivity model for single-case data.

When performing the two-sided RT in a between-subject design, we use the randomization schedule of a completely randomized design (CRD) (Onghena & Edgington, 2005). Suppose we want to test whether the treatment effect Δ equals 3.375. In this case, the model of unit-treatment additivity expresses the observed score of X_i^B as:

$$X_i^B = X_i^A + 3.375$$

This implies that we can test the null hypothesis (H_0) that $\Delta = 3.375$ against the alternative hypothesis (H_1) that $\Delta \neq 3.375$ by subtracting Δ from all X_i^B s in the data and then by performing the RT. Depending on whether the RT rejects H_0 at significance level α , we can determine whether or not 3.375 is in the $(100 - \alpha) \%$ CI. Constructing a CI for Δ through RTI consists of executing the RT for multiple absolute values of Δ and retaining all values that are not rejected by the RT. In this approach, each iteration of the search uses a two-sided RT at the α level in order to gain a $(100 - \alpha) \%$ CI.

When using RTI with a computer algorithm, the step size for changes in Δ from one iteration to the next must be determined in advance. The precision of the RTI procedure (and of the resulting CI) can be increased by lowering the step size, but a smaller step size will also require more computing time before the boundaries of the CI are reached. An efficient algorithm that balances precision and computing time uses a stepwise iterative procedure, which minimizes the number of iterations needed to calculate the boundaries of the CI for a certain amount of decimals. The algorithm we propose here calculates the boundaries of the CI at any number of decimals by using a stepwise iterative procedure. This procedure starts with a relatively large step size to get a rough estimate of the boundaries very quickly. Next, the algorithm starts again from the last Δ that yielded a p -value that is larger than the significance level, but this time with a step size that is ten times as small. This procedure is repeated until the algorithm arrives at the boundaries of the CI for the desired precision. In this way, the boundaries of the CI can be determined very accurately without requiring a lot of unnecessary iterations of the algorithm. Note that we will use two decimals to report CIs for the examples in this paper.

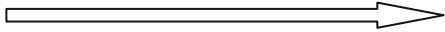
The search method for the chronic pain experiment with intermediate output of the algorithm with the step size set to 0.005 is illustrated in Table 3.

$$\text{Lower boundary} : 3.375 - 2.125 = 1.25$$

$$\text{Upper boundary} : 3.375 + 2.125 = 5.50$$

The 95 % CI for Δ for the chronic pain experiment is [1.25–5.50] and the two-sided p -value is .0014. Because RTI assumes a constant additive treatment effect, Δ can be interpreted as the difference in the average quality-of-life score between the patients of the control condition and the patients of the experimental condition. The interpretation of the nonparametric CI is similar to the parametric CI following the t -procedure, but the repetitions do not refer to repeated random sampling but to repeated random assignments of the patients to the two conditions. Note that $\Delta = 0$ is not included in the interval which confirms that the sample means differ from each other in a statistically significant way (see RT p -value of .0014 calculated previously). This illustrates that the CI conveys the same information as the hypothesis test, with the advantage of providing a range of “plausible values” for the computed ES (du Prel, Hommel, Röhrig, & Blettner, 2009). Also note that the abandonment of the implausible assumptions of random sampling and normality of the population distributions leads to a CI that is somewhat wider than the CI in the parametric case. This means that the broader applicability and general validity of nonparametric CIs comes at the expense of a small loss of statistical power, as compared to parametric CIs.

Table 3 Intermediate output of the randomization test inversion (RTI) algorithm for the hypothetical data of the chronic pain experiment using a two-sided randomization test. Note that $|\Delta - \theta_{obs}|$ is the RTI equivalent of θ_{diff} from Table 2



$ \Delta - \theta_{obs} $	0	...	2.120	 2.125 	2.130
Δ	3.375	...	5.495	5.50	5.505
p	10733	.0733	.0412
In CI	YES	...	YES	YES	NO

Nonparametric CI for a standardized mean difference

As a way of standardizing the mean difference statistic from the previous example, we can divide the mean difference between the A condition and the B condition by a standard deviation term such that we obtain a d statistic. Our d statistic then becomes:

$$d = \frac{\bar{B} - \bar{A}}{s_{A,B}}$$

Note that $s_{A,B}$ is the standard deviation of all observed data before Δ is subtracted from the B scores and is thus a constant for the calculation of the test statistic for each assignment and for each tested Δ . If $s_{A,B}$ would be calculated after Δ is subtracted from the data of the experimental condition, $s_{A,B}$ increases when Δ is increased and results in an inflation of the variance of the randomization distribution. In this case, different Δ s would be tested using randomization distributions with different variances. This is not desirable as the unit-treatment additivity model does not contain a variance parameter. In sum, by standardizing the mean phase difference by $s_{A,B}$ we can adjust for the overall variability of the observed data and simultaneously remain within the unit-treatment-additivity model as is the case in the example for the unstandardized mean difference. The observed d -value for the pain example is 1.5050 and the 95 % CI is [0.56–2.45]. The two-sided p -value equals .0014. Note that this p -value is exactly the same as the p -value for the unstandardized mean difference. This is because both test statistics differ only by a constant. As such, the position of the observed test statistic relative to the randomization distribution is exactly the same for the unstandardized mean difference as for the standardized mean difference which leads to identical two-sided p -values (Edgington & Onghena, 2007).

Nonparametric CIs for single-case designs

In the previous sections, we have introduced the use of RTI to derive a nonparametric CI for an unstandardized and standardized mean difference in a well known setting: the comparison

of two independent samples in a completely randomized between-subject design. We will now demonstrate how this method can be applied to derive RTI CIs for ES measures in SCEs, a setting in which the random sampling and distributional assumptions have traditionally been contested (Hartmann, 1974; Houle, 2009; Kratochwill et al., 1974). Note that in the case of an SCE, the repeated measurements are the experimental units whereas in a between-subject design the experimental units refer to the different subjects.

Various types of single-case designs can be broadly categorized into two groups: alternation designs and phase designs (Onghena & Edgington, 2005). In alternation designs, every measurement occasion can be randomly (or randomly within certain restrictions) assigned to every level of the independent variable. Examples of alternation designs include the completely randomized design (CRD), the alternating treatments design (ATD), and the randomized block design (RBD) (Barlow et al., 2009; Onghena & Edgington, 1994). Phase designs divide the sequence of measurement occasions into separate treatment phases, with each phase containing multiple measurements. Examples include the AB design and extensions thereof such as the ABA and the ABAB design (Kratochwill & Levin, 2014; Onghena, 1992).

Individual SCEs can be replicated using two types of strategies: simultaneous replication or sequential replication (Onghena & Edgington, 2005). Using the simultaneous replication strategy, multiple alternation or phase designs are executed at the same time. The best-known example of a simultaneous replication design is the multiple baseline across participants design (MBD), which combines multiple phase designs (usually multiple AB designs) and in which the treatment is administered in a time-staggered manner across the individual participants. In the sequential replication strategy, individual SCEs are replicated sequentially in order to test the generalizability of the effect to other participants, settings, or outcomes. For both replication strategies, a multivariate test statistic or p -value combining can be used to evaluate the null hypothesis of no treatment effect for any of the individual experiments (Bulté & Onghena, 2009; Koehler & Levin, 1998; Marascuilo & Busk, 1988; Onghena, 1992; Onghena & Edgington, 2005).

In the following sections, we will use hypothetical research examples to illustrate RTI for single-case alternation designs (CRD, RBD, and ATD) as well as for single-case phase designs (AB, ABA, and ABAB). Additionally, we will discuss how RTI can be extended to simultaneously and sequentially replicated designs.

In the hypothetical research examples, we will use the unstandardized and standardized mean difference as ES measures for the alternation designs and we will use the immediate treatment effect index (which is also a mean difference type test statistic) for the phase designs. However, we should point out that there is currently little consensus in the SCE community regarding which type of ESs are optimal for use with single-case data (e.g., Campbell & Herzinger, 2010; Lane & Gast, 2014). Ideally, different types of effects require different types of ESs which are optimally capable of capturing a specific type of effect. These effects can be visually explored, complementary to the statistical analysis, but whatever measure of ES is selected, valid statistical inference requires that the expected effects are specified beforehand. If one is interested in different types of effects beforehand (e.g., level, variability, trend, overlap, immediacy, and consistency) then a battery of tests or CIs can be administered simultaneously. In that case the familywise Type I error rate also has to be controlled accordingly (see Westfall & Young, 1993 for procedures that fit within the RT rationale). That being said, the mean difference type ESs that we will use in the next examples are most appropriate for research situations in which one wants to test or quantify the differences in level or the immediacy of effects. If other effects are expected then other ES measures should be put in place.

Nonparametric CIs for single-case alternation designs

Completely randomized single-case designs

The single-case design that is most similar to the completely randomized between-subject design is the completely randomized single-case design. The randomization scheme of both designs is identical, yielding the same collection of permissible assignments, with the only difference being that the experimental units in the single-case CRD are the repeated measurements of the same subject whereas the experimental units in the between-subject CRD are measurements from different subjects.

Consider the following hypothetical SCE: A therapist wants to evaluate a customized psychotherapy treatment, designed to elevate the self-esteem of a patient suffering from depression. The SCE lasts 10 days, where five days are randomly selected for administering the treatment. No treatment is administered on the remaining five days. At the end of each day, the patient reports his perceived self-esteem, expressed on a scale from 1 to 10. The randomization of the treatment condition (B) and the control condition (A) to the measurement occasions is done by choosing one of the permissible assignments generated by the CRD randomization scheme. Suppose we randomly select the following assignment: ABAABBABAB. Hypothetical data for this SCE are shown in Fig. 1.

The mean self-esteem scores are 3.8 and 6.2 for A and B, respectively, and the observed difference in means ($\theta_{obs} = \bar{B} - \bar{A}$) equals 2.4. Application of exactly the same algorithm as presented before (RTI), results in a 95 % CI of [0.40–4.40].

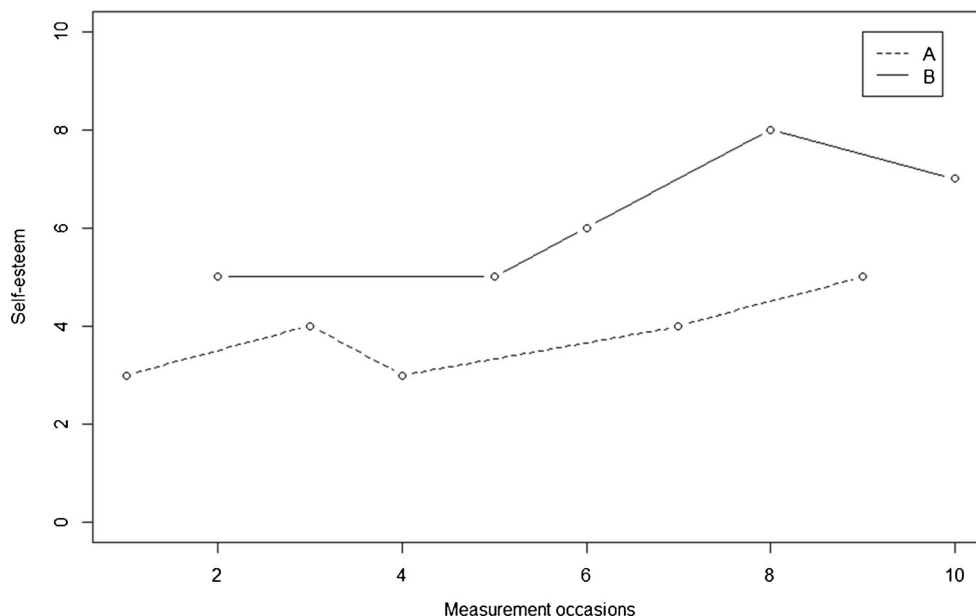


Fig. 1 Hypothetical data from a completely randomized single-case design evaluating the effect of psychotherapy on self-esteem in a single depressed patient on a 1–10 rating scale

For the standardized mean difference, the observed value is 1.4670 with a 95 % CI of [0.25–2.69]. In both cases, the two-sided p -value is .0238. Note that the value 0 is not included in any of the intervals, suggesting a statistically significant treatment effect on the perceived self-esteem of the depressed patient.

Randomized block single-case designs

Building further on the previous example, suppose the same therapist now wants to evaluate another customized treatment for another depressed patient with the difference being that this patient typically has good and bad days influencing the patient's perceived self-esteem, irrespective of the treatment administered on those days. In this case, the therapist can use a randomized block design for his SCE to control for this confounding factor. Suppose the experiment consists of 10 days (i.e., 10 blocks) where on each day, the therapist administers both treatments (i.e., the control condition and the actual treatment condition) and records the perceived self-esteem of the patient after administering each treatment (i.e., two self-esteem scores per day). The sequence of conditions within a day (i.e., block) is determined randomly for every day. Assume the therapist randomly selects the following assignment to collect the data: AB | BA | BA | AB | BA | BA | AB | BA | BA | AB. The hypothetical data for this SCE are presented in Fig. 2.

For this example, the mean self-esteem scores are 5.5 and 4.9 for A and B, respectively, and the observed difference in means ($\theta_{obs} = \bar{B} - \bar{A}$) is -0.6 . The 95 % CI is $[-1.60-0.40]$. The observed value for the standardized mean difference is

-0.5430 with a 95 % CI of $[-1.45-0.36]$. Both for the unstandardized mean difference and standardized mean difference, the two-sided p -value is 0.390625. Because the value 0 is included in both intervals and because the two-sided p -values are larger than any conventional significance level, we can conclude that the unstandardized and standardized treatment effect for this particular patient are not statistically significant.

Alternating treatments single-case designs

Finally, the therapist can also use an alternating treatments design (ATD) to evaluate the treatment effect of the customized psychotherapy treatment on a single patient. Such a design can be used when rapid alternation between the control condition and the treatment condition is required. The randomization scheme of an ATD is similar to that of a CRD but assignments with a predefined number of consecutive administrations of the same condition are excluded (Onghena & Edgington, 1994). Suppose the experiment consists of ten measurement occasions: Five baseline observations and five treatment observations and requires that no more than three subsequent measurement occasions can belong to the same condition. A permissible assignment for the experiment would then be: AABABAABBB. Figure 3 illustrates a hypothetical dataset for this design.

For this example, the mean phase scores are 6.8 and 7 for A and B, respectively, and the observed difference in means ($\theta_{obs} = \bar{B} - \bar{A}$) is 0.2. The 95 % CI is $[-0.80-1.20]$. The observed value for the standardized mean difference statistic is 0.2011 with a 95 % CI of $[-0.80-1.21]$. In both cases, the two-sided p -value is 1. Note again that the value 0 is included in both

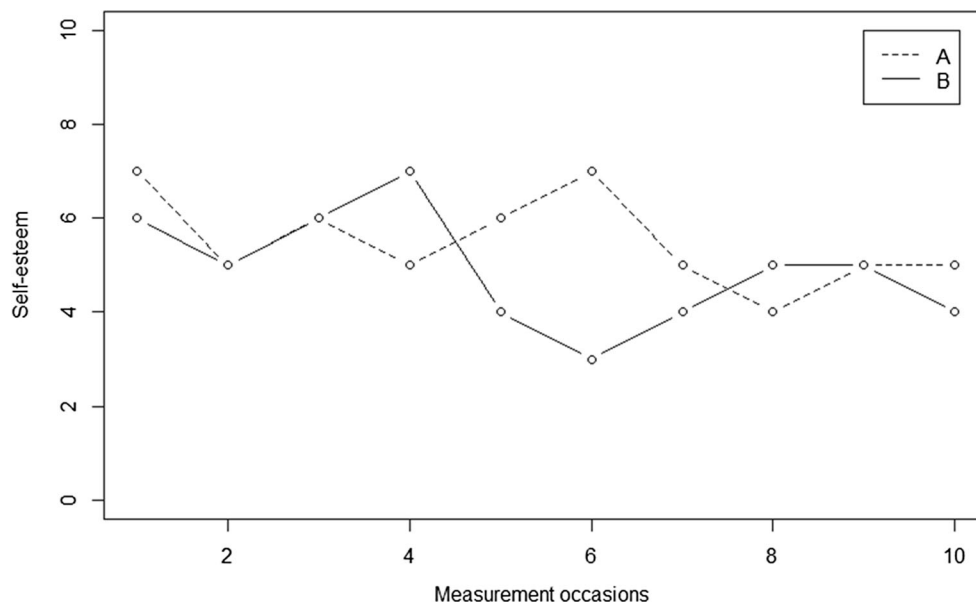


Fig. 2 Hypothetical data from a single-case randomized block design evaluating the effect of psychotherapy on self-esteem in a single depressed patient on a 1–10 rating scale

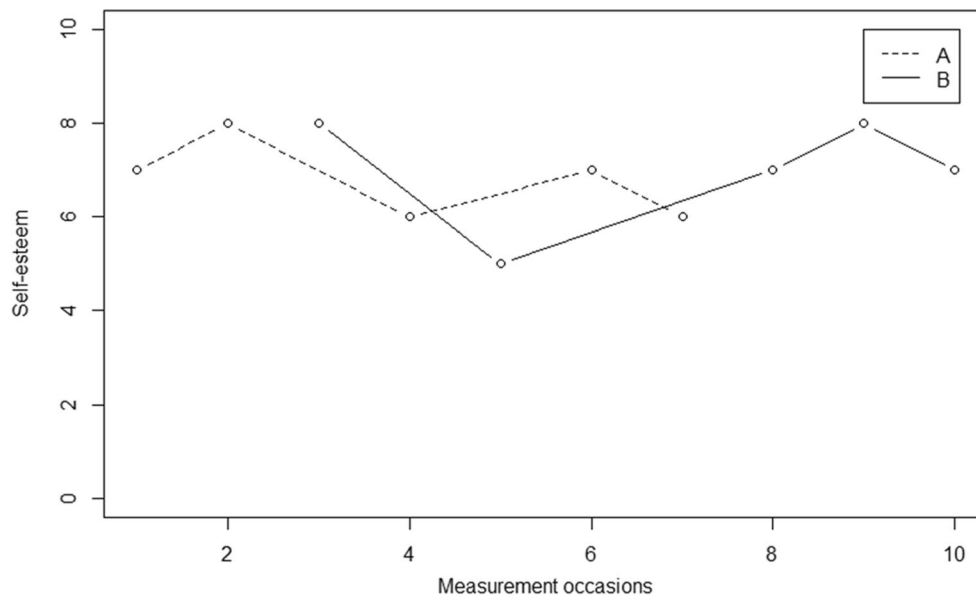


Fig. 3 Hypothetical data from a single-case randomized alternating treatments design evaluating the effect of psychotherapy on self-esteem in a single depressed patient on a 1–10 rating scale

intervals and that both two-sided p -values have attained their maximum value of 1, evidently indicating that there is no significant treatment effect.

In this section, we illustrated RTI for three different single-case alternation designs (CRD, RBD, and ATD). Note that the RTI method is essentially the same for the three illustrated designs with the sole difference that the employed randomization scheme of the RT is specific to each of the three designs.

Nonparametric CIs for single-case phase designs

In the next section, we will focus on single-case phase designs. To recapitulate, phase designs divide the sequence of measurement occasions into separate treatment phases with each phase containing multiple measurements. As a consequence, randomization of the condition labels in an RT for phase designs can only pertain to the moment of phase change but not to the treatment order within phases. In the following section, we will illustrate RTI for an AB, ABA, and ABAB phase design using three different examples.

AB phase designs

The simplest example of a phase design is an AB design. This design consists of a certain number of baseline phase observations followed by a certain number of treatment phase observations. As mentioned before, the specific assignments that are permissible for the RT depend on the characteristics of the design (e.g., Onghena & Edgington, 2005). In an AB phase design, all A observations precede all B observations. Consequently, the

randomization of the condition labels can only pertain to the available moments of phase change. The number of moments that qualify as potential moments of phase change is determined by the required minimum phase length, which must be predefined by the experimenter. For example, the experimenter could require that for an SCE with 20 measurement occasions, the baseline phase and treatment phase should at least have four observations in each permissible assignment. For an AB phase design in general, the number of data randomizations for N observations with a minimum phase length of k is equal to $N - 2k + 1$ (Onghena, 1992).

Suppose a therapist wants to evaluate the effectiveness of a customized behavioral treatment to reduce anxiety (measured on a scale from 0 to 15) in a patient with post-traumatic stress disorder. In order to do so, the therapist conducts an SCE consisting of 69 measurement occasions with a minimum phase length requirement of eight observations. For such a design, the number of permissible assignments equals 52 and the smallest attainable p -value is about .02 (1/52). Suppose the therapist randomly selects an assignment consisting of 32 baseline phase observations and subsequently 37 treatment phase observations.

The hypothetical scores for the baseline phase (denoted by A) and treatment phase (denoted by B) are displayed in Fig. 4.

An important indicator for treatment effectiveness in AB designs is the immediacy of the treatment effect (Kratochwill et al., 2010). Let us illustrate the RTI for the *immediate treatment effect index* (ITEI). Following the recommendation by Kratochwill et al. (2010) we define the ITEI in an AB phase design as the average difference between the last three A

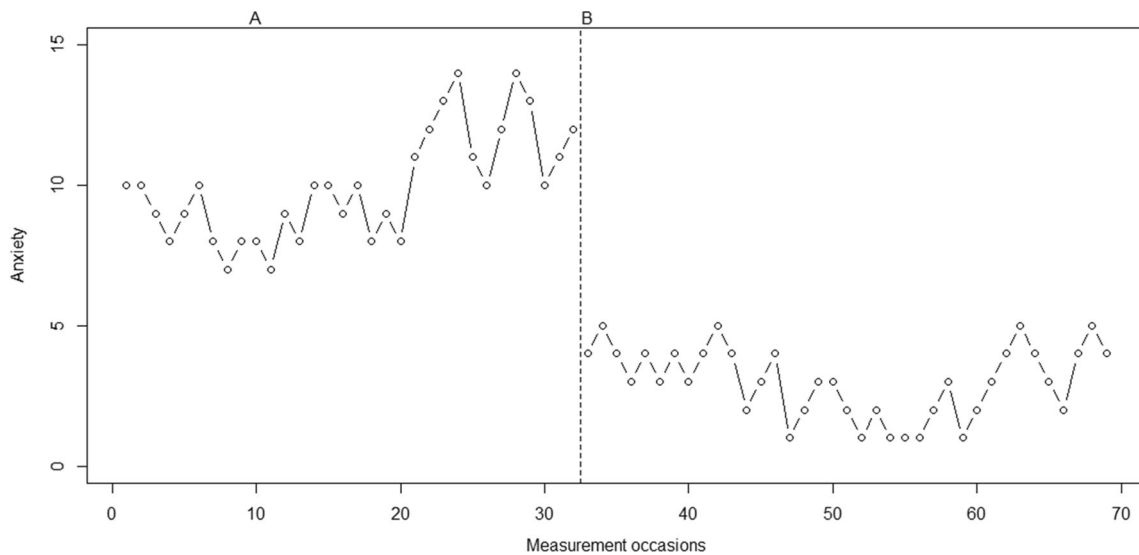


Fig. 4 The hypothetical scores of an AB phase design evaluating the effect of a customized behavioral treatment on anxiety in a patient with post-traumatic stress disorder on a 1–15 rating scale

observations and the first three B observations. Denote the average of the last three A observations as \bar{A}_{ITEI} and the last three B observations as \bar{B}_{ITEI} . For this example, we choose to calculate $\bar{A}_{ITEI} - \bar{B}_{ITEI}$. The observed value is 6.6667 and the 95 % CI using the RTI method turns out to be [2.33–11.00]. The two-sided *p*-value for the observed ITEI equals .01852. The *p*-value and the 95 % CI indicate that the patient showed significantly lower anxiety levels during the first three measurement occasions in the treatment phase compared to the last three measurement occasions of the baseline phase.

ABA phase designs

An ABA phase design or withdrawal design is an extension of a basic AB phase design where the treatment phase is followed by a return to the baseline phase. Returning to our example of a customized behavioral treatment to reduce anxiety in a post-traumatic stress patient, consider an ABA phase design with 38 measurement occasions and a minimum phase length of six measurement occasions in order to evaluate treatment effectiveness. Suppose the therapist randomly selects the assignment which comprises ten baseline phase observations followed by 20 treatment phase observations and then again followed by eight baseline phase observations. This assignment can be graphically expressed as:

AAAAAAAAAABBBBBBBBBBBBBBBBBB
 BBBBAAAAAAAA

Figure 5 shows a hypothetical dataset for this design.

In contrast to an AB phase design, an ABA phase design has two distinct moments of phase change. Since

both moments offer the possibility to observe a potential treatment effect, the ITEI must be calculated from a score range that includes both phase changes. For this example, we again choose to subtract the B observations from the A observations. More specifically, the ITEI is calculated by taking the mean difference of the pooled B observations and the pooled A observations that surround each moment of phase change:

AAAAAAAAA₁₁A₁₂A₁₃B₁₁B₁₂B₁₃BBBBBBBBBBBBB
 BBB₂₁B₂₂B₂₃A₂₁A₂₂A₂₃AAAAA

The ITEI then equals (in case when the B averages are subtracted from the A averages):

$$\left(\frac{\sum_{i=1}^2 \sum_{j=1}^3 B_{ij}}{6} - \frac{\sum_{i=1}^2 \sum_{j=1}^3 A_{ij}}{6} \right)$$

Note that the way in which we calculate the immediate treatment effect in this example is not a universally accepted conceptualization of a treatment effect in an ABA phase design. More specifically, it can be argued that the introduction of a treatment (i.e., change from A phase to B phase) is a qualitatively different effect than the removal of a treatment (i.e., change from B phase to the second A phase). In this view, the A and B scores from both moments of phase change should not be pooled together as they represent different effects. Single-case researchers that want to look at only one moment of phase change can do so by simply taking only the data from two phases and analyze them as if they were either an AB or BA phase design (cf. the hypothetical example for an AB phase design).

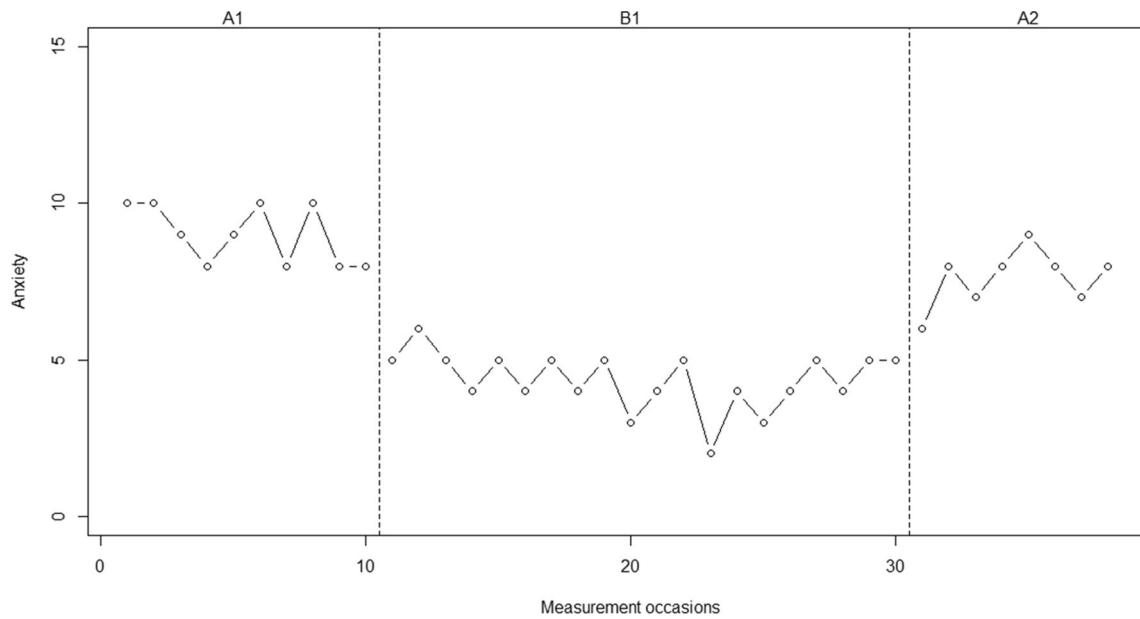


Fig. 5 The hypothetical scores of an ABA phase design evaluating the effect of a customized behavioral treatment on anxiety in a patient with post-traumatic stress disorder on a 1–15 rating scale

For the current example, the observed ITEI value is 2.83 and the 95 % CI for the ITEI is [0.67–5.00]. The two-sided *p*-value equals 0.0087. The *p*-value and the 95 % CI suggest a significant average difference in the average anxiety score between the last three A observations and the first three B observations for both moments of phase change.

AAAAAAAAA₁₁A₁₂A₁₃B₁₁B₁₂B₁₃BBBBBBB₂₁B₂₂
 B₂₃A₂₁A₂₂A₂₃AAAAA₃₁A₃₂A₃₃B₃₁
 B₃₂B₃₃BBBBBBBBB

The ITEI then equals (in case when the B averages are subtracted from the A averages):

$$\left(\frac{\sum_{i=1}^3 \sum_{j=1}^3 B_{ij}}{9} - \frac{\sum_{i=1}^3 \sum_{j=1}^3 A_{ij}}{9} \right)$$

Note that the remark we made about the ITEI for ABA designs regarding the possible qualitative difference between introducing a treatment and removing a treatment also applies in the case of an ABAB design. For this example, the observed value of the ITEI is 2.78 and the 95 % CI is [0.22–5.33]. The two-sided *p*-value is 0.0309. The *p*-value and the 95 % CI indicate that the average difference in the average anxiety score between the last three A observations and the first three B observations for three moments of phase change is significant.

ABAB phase designs

Alternatively, the therapist from the previous example could also have used an ABAB phase design to evaluate treatment effectiveness. This design is essentially a double AB design where an initial first baseline phase and treatment phase are followed by a second baseline and treatment phase. Suppose the therapist chose an ABAB phase design with a total of 44 measurement occasions and a minimum phase length of four observations. Assume the therapist randomly selects an assignment which consists of ten A phase observations and 12 B phase observations for the initial A phase and the initial B phase, followed by ten A observations and 12 B phase observations for the second A phase and the second B phase, respectively. This assignment can be graphically expressed as:

AAAAAAAAAABBBBBBBBBBBBAAAAAAAAA
 AABBBBBBBBBBBB

Figure 6 displays a hypothetical dataset for this design. Using a similar approach as in the ABA phase design, the ITEI can again be calculated from the group of pooled A scores and pooled B scores surrounding every moment of phase change:

Nonparametric CI for replicated single-case designs

RTI can also be applied to simultaneously and sequentially replicated single-case designs. Consider for example a situation in which a researcher wants to obtain a nonparametric CI for the treatment effect in an MBD or for the average treatment effect for a group of sequentially replicated AB designs. Bulté and Onghena (2009) describe an RT for the MBD which can

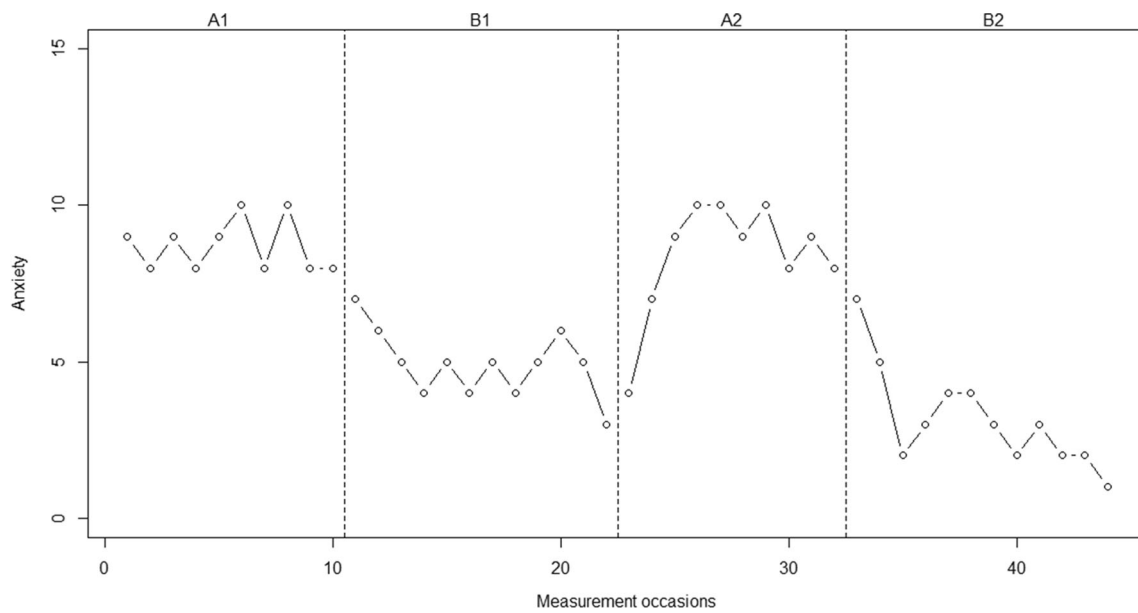


Fig. 6 The hypothetical scores of an ABAB phase design evaluating the effect of a customized behavioral treatment on anxiety in a patient with post-traumatic stress disorder on a 1–15 rating scale

be equally well used for sequentially replicated single-case designs. In this RT, the assignments are constructed by randomizing the possible start points of the intervention in each AB design as well as randomly assigning the subjects to the different AB designs. Next, the chosen test statistic (e.g., the ITEI) is averaged across all the individual AB designs and a single reference distribution for this multivariate test statistic is constructed by repeating the calculation for all assignments. The p -value of this test statistic can be calculated by comparing the observed value of the test statistic to the reference distribution. This RT can then be inverted through RTI in order to obtain a nonparametric CI for the selected ES measure. Note that in this context, the CI is a measure of uncertainty for the average treatment effect across all individual SCEs that were included in the analysis.

Discussion

The main purpose of this paper was to propose an RTI method to construct nonparametric CIs for ES measures in a variety of single-case designs. Starting with a between-subject situation, we illustrated how the inversion of a t -test can provide a parametric CI for the difference in means between two independent samples. We then proposed the inversion of an RT to obtain a nonparametric CI when making classical parametric assumptions is not desirable. We provided examples of RTI for a variety of single-case alternation and phase designs such as CRD, ATD, RBD, AB, ABA, and ABAB designs, and we also explained how RTI can be used for simultaneous and sequential replication designs. In this section, we will

make a few more remarks regarding the proper use of the RTI method.

First of all it should be emphasized that the RT and RTI method only have guaranteed validity if some form of random assignment has taken place (i.e., the random assignment assumption of the RT). Evidently, without random sampling or random assignment any inference is purely observational or based on assumed random processes. That being said, single-case researchers can benefit greatly from incorporating random assignment into their designs as randomization increases both the internal validity and the statistical conclusion validity of SCEs (e.g., Cook & Campbell, 1979; Edgington & Onghena, 2007; Kratochwill & Levin, 2010; Shadish, Cook, & Campbell, 2002). Edgington (1996) argues that by incorporating randomization, single-case studies become experimental studies from which valid causal inferences can be made by means of RTs. Randomization strengthens the SCE's internal validity because it yields statistical control over known and unknown confounding variables (Levin & Wampold, 1999; Onghena, 2005). Confounding variables such as serial correlation, history effects, and maturation effects may have an influence on the observed data, but all these potential effects are constant under the permissible assignments of the randomization scheme if the null hypothesis is true. As such, these potential effects cannot be considered as explanations in case a significant treatment effect is found. In addition, randomization also increases the statistical conclusion validity of the SCE because it leads to a statistical test that is based on the randomization as it actually took place in the executed design (i.e., the RT) (Kratochwill & Levin, 2010; Onghena & Edgington, 2005).

Secondly, readers should realize that RTI additionally requires a treatment effect model. In this manuscript we have selected the unit-treatment additivity model because it is the simplest model and because it is most popular and well studied within nonparametric statistics (e.g., Cox & Reid, 2000; Hinkelmann & Kempthorne, 2008; Lehman, 1959; Welch & Gutierrez, 1988). This model assumes that the scores of the experimental units in the baseline condition and the treatment condition differ only by a constant additive treatment effect (Δ). The goal of RTI is then to determine a CI for Δ . Note that this assumption is not inherent to RTI but rather to the model that is chosen to reconstruct the null scores from the observed data. In contrast, a nonparametric p -value requires no such assumption but also offers less information than a CI. In this sense, the construction of a nonparametric CI yields more information than a nonparametric p -value but comes at the cost of an extra assumption regarding the model that provides an accurate description of the observed data. This demonstrates an intriguing general rule: CIs require an additional assumption (viz., the hypothetical effect function), as compared to a bare-bones significance test. As such the advocacy for the “new statistics” (Cumming, 2012) in which significance tests are replaced by effect sizes and CIs is far too simplistic. Also in parametric statistics there is an implicit assumption that the treatment effect is best represented by a mean shift if a CI for a difference between means is chosen, an assumption which is similar to the unit-treatment additivity model of nonparametric statistics.

An important question then is whether the unit-treatment additivity model provides an accurate description of the effect in single-case data. Of course, there is no general answer to this question because of the idiosyncratic nature of any dataset. However, single-case data comprises repeated measurements of the same person. For this reason, the treatment effect may be influenced by 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; Parker, Cryer & Byrns, 2006). Furthermore, the onset of the treatment may also interact with these time-related effects (e.g., a change in trend after the onset of the treatment compared to the baseline phase, see Van den Noortgate & Onghena, 2003, or a change in variability after the onset of the treatment, see Ferron, Moeyaert, Van den Noortgate, & Beretvas, 2014). These types of effects are not accounted for in the unit-treatment additivity model and as such may confound treatment effect estimates when they are present in the analyzed SCE data.

In addition, because the unit-treatment additivity model only contains a parameter that represents the treatment effect as one overall difference, it is most sensible to calculate nonparametric CIs for mean difference type ESs such as the ones we used in the illustrations in this article. For example, it would not be meaningful to use an ES that measures

variability because the unit-treatment additivity model does not contain a parameter that accounts for variability in the data. Summarized, the RTI method using the unit-treatment additivity model is most adequate to be used with mean difference type ESs and for datasets with small or no trends and with relatively low variability.

However, as we discussed above, it is important to realize that the aforementioned limitations are limitations of the unit-treatment additivity model, not of RTI itself. In fact, a major strength of RTI is that one can use models other than the unit-treatment additivity model in order to reconstruct the null scores from the observed scores. More specifically, all statistical models for null hypothesis testing assume some sort of relation between the vector of null scores $X_1^A, X_2^A, \dots, X_i^A$ and the vector of observed scores $X_1^B, X_2^B, \dots, X_i^B$ such that for any null score X_i^A and its observed score X_i^B the following equation holds:

$$X_i^B = f(X_i^A)$$

with f being a generic effect function. In the case of the unit-treatment additivity model, $f(x)$ equals $x + \Delta$. A slightly more flexible model is the extended unit-treatment additivity model in which $f(x)$ equals $x + \Delta + \varepsilon_i$ where the ε_i are independent and identically distributed random variables with a mean of zero (Cox & Reid, 2000). By including the ε_i random variables, the treatment effect can vary between experimental units. Alternatively, one can also formulate a model that contains a trend component that can be used for datasets with deterministic trends: $f(x) = x + \Delta + t\beta$ with t indicating the number of the measurement occasion in the treatment phase and β being a constant trend effect. By including a time variable t into the model, one could also account for delayed treatment effects. The aforementioned models are all examples of additive models. In contrast to additive models, multiplicative models assume a nonlinear relation between the null scores and the observed scores. An example would be $x + \frac{1}{x}\Delta$ in which the magnitude of the treatment effect for experimental unit i is inversely related to its null score. This means that the treatment effect is smaller for large null scores than for small null scores.

In sum, although the unit-treatment additivity model might not be a perfect model for all types of single-case data, the examples above illustrate that various types of effects can be included into the model that is used for RTI. While $f(x)$ can in principle be any function, it is important to keep in mind that the chosen statistical model regarding the treatment effect must be plausible and well interpretable. In this respect, the unit-treatment additivity model that we adopt here is a generally accepted statistical model for performing nonparametric tests and constructing nonparametric CIs (e.g., Cox & Reid, 2000; Garthwaite, 2005; Hinkelmann, & Kempthorne, 2008; Lehman, 1959; Welch & Gutierrez, 1988). Future research

should focus on comparing other models with the unit-treatment additivity model with respect to modeling SCE data for use within RTI and investigating the effect on the resulting nonparametric CIs in case the employed RTI model is not adequate for the observed data.

We already mentioned that the RTI approach assumes a random assignment model instead of a random sampling model. This is an important distinction with respect to the nature of the statistical inferences resulting from RTI. Whereas adopting a random sampling model (which is done in most parametric tests) aims at making inferences regarding the population values of certain parameters, the random assignment model only allows us to make causal inferences regarding the observed data (e.g., Ernst, 2004; LaFleur & Greevy, 2009; Ludbrook & Dudley, 1998). This means that when one constructs a 95 % CI for the ITEI in a single AB phase design, conducted to evaluate a chronic pain treatment for a certain patient for example, the CI does not apply to other patients which have received the same chronic pain treatment.

Although there are numerous advantages of using RTs instead of parametric tests for the analysis of single-case data, there are also some limitations, which have been recently outlined by Heyvaert and Onghena (2014). Because RTI is based on the equivalence between the RT and the CI that results from its inversion, it is sensible to briefly review these limitations as they equally apply for the type of CIs we propose here.

One limitation that is particularly relevant for RTI concerns statistical power. As mentioned before, one of the factors that contribute to the statistical power of the RT is the number of permissible assignments for a given experimental design. Whereas the randomization schemes of alternation designs almost always yield sufficient assignments to allow adequate statistical power, the number of permissible assignments in an AB phase design with relatively few measurement occasions and/or a large minimum phase length can be quite small which compromises the statistical power of the RT. More specifically, if a given AB phase design has less than 20 permissible assignments, it is not possible for the RT to achieve a p -value smaller than .05 and consequently it is not possible to construct a 95 % CI via RTI (although an interval with a smaller confidence level is possible). However, it should also be taken into account that an AB phase design is the “weakest” single-case phase design in this respect: For more advanced phase designs such as the ABA and ABAB phase designs, the potential issue of insufficient permissible assignments is generally lacking. For example, whereas an AB phase design with 24 observations and a minimum phase length of two observations yields 21 permissible assignments, an ABAB design with the same number of observations and minimum phase lengths yields 969 permissible assignments which is more than enough to achieve small p -values.

Note also that the AB and ABA phase designs do not meet the What Works Clearinghouse single-case design standards (Kratochwill et al., 2010). In this sense, the results of the statistical power of AB and ABA phase designs confirm the judiciousness of these standards for phase designs with a small number of observations. In order to meet these standards, an SCE must have a minimum of four distinct phases (such that there are at least three moments where a potential treatment effect can be observed) with at least five data points per phase (if the SCE has three data points per phase it meets standards with reservations) (Kratochwill et al., 2010). In addition, Onghena and Edgington (2005) suggest that the weak statistical power of AB designs can be boosted considerably when three or more AB designs are combined in a multiple-baseline design or in a series of sequential AB designs, a suggestion that was recently supported in an extensive simulation study by Heyvaert et al. (2016).

Finally, a remark concerning the proposed ITEI for phase designs is in order. A disadvantage of the ITEI that we used as an ES for phase designs might be that this statistic is calculated based on only three data points for each phase. If the ITEI is to capture a potential treatment effect, the effect must be immediately and effectively observable when comparing the last three A observations and the first three B observations (surrounding the moment of phase change). Some treatment effects, however, might not be immediately observable (e.g., a delayed effect or an effect that occurs gradually over time) and thus will not be captured well by the ITEI. However, given that phase designs feature relatively long sequences of serially correlated observations in the same experimental condition, a gradual treatment effect is hard to distinguish from trend effects that are independent of the experimental condition. For this reason, an immediately observable score change in a phase design is a strong indicator for a potential treatment effect. Alternatively, it is possible to calculate the ITEI from a larger range of values surrounding the moment of phase change such that more gradual treatment effects could also be detected.

Future research directions: Developing nonparametric CIs for nonoverlap statistics

An important advantage of the RT is that it can be used with any type of ES as the test statistic (Heyvaert & Onghena, 2014). Therefore, RTI can be employed to obtain nonparametric CIs for any ES provided that the model that generates the scores under the alternative hypothesis contains a parameter that pertains to the type of effect that the ES is sensitive for. In this article, we used RTI with the unit-treatment additivity model and mean difference type ESs. The unit-treatment additivity model can also be used for ESs that are not mean differences but do measure mean level separation. The only

Table 4 Description of arguments for the randomization test inversion functions. Each row in the table contains the name of a specific function argument, along with a short description and possible values

<i>Data</i>	A data frame that contains the measurements. It must consist of two columns (separated by a tab) where the first column contains the condition labels and the second column contains the recorded scores. Each row of the data frame must contain only one measurement occasion
<i>Design</i>	Specifies the experimental design of the SCE. Possible values are “CRD”, “ATD”, “RBD”, “AB”, “ABA” and “ABAB”
<i>Direction</i>	Specifies whether for the calculation of the mean difference, standardized mean difference or ITEI the function should take $\bar{A} - \bar{B}$ or $\bar{B} - \bar{A}$ and thus accepts the values “A-B” and “B-A” respectively
<i>Limit</i>	Signifies the minimum phase length when the experimental design is an AB phase design and the maximum number of consecutive measurement occasions of the same condition when the design is an ATD
<i>Number (only for random versions)</i>	Signifies the size of the random sample of permissible assignments. The default value is 2000
<i>Range</i>	This argument can be specified if the user wishes to calculate the ITEI in phase designs from a larger range of values (the default value is three). The “range” value equals the amount of observations per phase that are incorporated in the ITEI calculation. For example, when “range” equals four in an AB, the ITEI is calculated from the four last A observations and the four first B observations. When “range” is set to zero in the function for the unstandardized mean difference in the context of a phase design, the automatic ITEI selection is overridden and the unstandardized mean difference is used as the test statistic
<i>Precision</i>	Specifies the number of decimals at which the CI will be calculated. The default value is two. Note that the algorithm will calculate the CI for “the requested amount of decimals + 1” and then round the boundaries to the requested amount of decimals. In this way, rounding errors are avoided
<i>Confidence</i>	Specifies the desired confidence level of the CI and accepts values from 0 to 1; a value of 0.95 is default.

SCE single-case experiment, *ITEI* immediate treatment effect index, *ATD* alternating treatments designs, *CI* confidence interval, *CRD* completely randomized single-case designs, *RBD* randomized block designs, *AB*, *ABA*, *ABAB* phase designs

requirement is that the respective ES has a monotonic relation with the constant additive shift parameter (Δ) of the unit-treatment additivity model. For this reason it would be interesting for future research to investigate the possibility of using nonoverlap statistics as ESs in RTI because these statistics measure mean level separation on an ordinal level.

Nonoverlap statistics are a group of nonparametric indices that are receiving considerable attention from the scientific community of single-case researchers. These nonoverlap statistics quantify the extent to which baseline phase data and treatment phase data do not overlap. This approach is rooted in the tradition of visually analyzing SCE data where data nonoverlap has been widely accepted as an indicator of treatment effectiveness (Sidman, 1960). An advantage of nonoverlap ESs is that they consider the individual data points in pairwise comparisons across phases and are therefore more robust to outliers in the data in comparison to mean level comparisons (Parker, Vannest, & Davis, 2011). Primarily in the last decade, a whole range of nonoverlap indices has been developed, including Percentage of Nonoverlapping Data (PND, Scruggs, Mastropieri, & Casto, 1987), Percentage of All Nonoverlapping Data (PAND, Parker et al., 2007), Improvement Rate Difference (IRD, Parker et al., 2009), Nonoverlap of All Pairs (NAP, Parker & Vannest, 2009), and Tau-U (Parker et al., 2011). Some of these ESs have no known sampling distribution (e.g., PND, PAND) and, as such, CIs for these measures cannot be constructed analytically. Other measures have a relationship with established statistical tests (e.g., NAP, Tau-U). For example, NAP is equivalent to

the Mann-Whitney U statistic (Mann & Whitney, 1947) and Tau-U is essentially a Kendall rank correlation (Kendall, 1938). Existing methods for constructing CIs for these measures are also based on permutation techniques, but all assume a completely randomized design (e.g., Bauer, 1972 for the Mann-Whitney U and Long & Cliff, 1997 for Kendall’s Tau).

As mentioned before, a major strength of the RT is the enormous flexibility with regard to the choice of the test statistic, and so also with regard to the choice of the ES measure that can be used as a test statistic (Heyvaert & Onghena, 2014). All previously mentioned nonoverlap indices could in principle serve as ES measures in RTI using the treatment-additivity model provided that they are sensitive for the mean level separation of two phases. In this way, nonparametric CIs can be calculated for these measures and, importantly, according to a randomization scheme that is in accordance with the used experimental design. Heyvaert and Onghena (2014) have already used an RT to calculate a nonparametric *p*-value for PND. An interesting avenue for further research would be to employ RTI using the unit-treatment additivity model to derive nonparametric CIs for nonoverlap statistics in randomized block designs, alternation designs, and phase designs.

Software availability

We have developed a set of easy-to-use R-functions to compute nonparametric CIs according to the method described in this paper. The R scripts for each of these

Table 5 Estimated runtimes for systematic randomization test inversions for alternation designs and phase designs. Cells with a dash indicate insufficient computer memory for that combination of sample size and experimental design

N	Alternation designs			N	Phase designs		
	CRD	ATD	RBD		AB	ABA	ABAB
10	< 1s	< 1s	< 1s	20	< 1s	< 1s	< 1s
12	< 1s	< 1s	< 1s	40	< 1s	< 1s	11s
14	1.3s	1.6s	< 1s	60	< 1s	1.5s	1m47s
16	4s	5s	< 1s	80	< 1s	4s	9m05s
18	10s	20s	< 1s	100	< 1s	7s	/
20	37s	2m16s	< 1s	200	< 1s	52s	/
22	/	/	< 1s	500	< 1s	/	/
24	/	/	3s				
26	/	/	4s				
28	/	/	9s				
30	/	/	12s				
...	/	/	/				

CRD completely randomized single-case designs, ATD alternating treatments designs, RBD randomized block designs

functions can be downloaded from <http://ppw.kuleuven.be/home/english/research/mesrg/appletsandsoftware>. Note that there are two ways in which an RT can be executed: (1) the systematic RT uses all of the permissible assignments to construct the randomization distribution, and (2) the Monte Carlo RT uses a random sample (of a pre-specified size) of the permissible assignments to construct the randomization distribution (Besag & Diggle, 1977).

In situations where the number of permissible assignments is very large (e.g., in the case of a CRD with a large number of measurement occasions), the Monte Carlo RT is computationally more efficient. It has been shown that the Monte Carlo RT produces valid *p*-values (Edgington & Onghena, 2007). In addition, the accuracy of the RT can be increased to the desired level simply by increasing the number of randomizations (Senchaudhuri, Mehta, & Patel, 1995). We have provided separate functions for the systematic and Monte Carlo versions of the RTI code. With regard to choice of test statistic, we have also provided separate functions for the unstandardized mean difference, the proposed *d* statistic. Note that the ITEI is selected automatically when the function for the unstandardized mean difference is used for phase designs (although this behavior can be overridden). The R-functions support all single-case designs for which we provided an example in the paper (i.e., CRD, RBD, ATD, AB, ABA, and ABAB). With respect to all of the supported randomization schemes, we have adapted code from the R package of Bulté and Onghena (2008) with permission from the authors. The four separate R functions are called: “mean_diff.systematic.randomization.ci()” for the systematic version using the mean difference test statistic, “mean_diff.random.randomization.ci()” for the random version using the mean difference test statistic, “d.systematic.randomization.ci()” for the systematic version using the *d* test statistic, and “d.random.randomization.ci()” for the random version using the *d* test statistic. Table 4 contains the various arguments of the functions along with all accepted values for these arguments.

An example of code that needs to be written in the R console to execute the function can be found below:

```
mean_diff.systematic.randomization.ci(data = AB_data, design = "AB", direction = "A-B", limit = 2)
```

This line of code executes systematic RTI in the context of an AB phase design with a minimal phase length of two observations and the ITEI as the

employed test statistic. The example code will construct a CI with a 95 % confidence level that is precise up to two decimals.

Table 6 Estimated runtimes for random randomization test inversions for alternation designs and phase designs using either 2,000 or 10,000 random samples. Cells with a dash indicate insufficient computer memory for that combination of sample size and experimental design

N	Alternation designs						N	Phase designs					
	CRD		ATD		RBD			AB		ABA		ABAB	
	R=2000	R=10000	R=2000	R=10000	R=2000	R=10000		R=2000	R=10000	R=2000	R=10000	R=2000	R=10000
50	5s	19s	26s	1m31s	7s	41s	100	3s	10s	3s	38s	20s	53s
100	6s	23s	29s	1m45s	12s	1m11s	200	3s	11s	10s	1m49s	3m48s	9m31s
150	8s	30s	1m5s	3m35s	15s	1m27s	500	4s	13s	7m	63m	/	/
200	13s	39s	1m31s	4m58s	22s	1m51s							

Two additional remarks regarding the RTI R-functions are in order. First, we recommend the use of systematic RTI whenever possible because it is the most precise procedure. In case the user's dataset is so large that the computer's processing power is insufficient to calculate the CI via the systematic way within a reasonable runtime, we recommend switching to random RTI. Random RTI may produce slightly different p -values and CIs each time the analysis is performed for the same dataset because of the random sampling of assignments. In contrast, the results from systematic RTI are always the same when the analysis is repeated for the same dataset because all of the permissible assignments are used to construct the randomization distribution.

Second, it should be noted that the runtime is dependent on the selected design, the sample size and the requested amount of decimals of the CI. For the sake of convenience for the user, we have provided a table with runtime estimates for various experimental designs and sample sizes. In order to obtain these runtime estimates, we used a Dell Optiplex 7010 computer with an Intel Core i5-3570 CPU (3.4 Ghz) and 4 GB of RAM memory, running on Windows 7 Enterprise 64 Bit. Note that these estimates may vary depending on the speed of the computer that is used. The estimated runtimes for systematic RTI can be found in Table 5 and the estimated runtimes for random RTI (using either 2000 or 10000 random samples) can be found in Table 6. For each cell in the table we generated two data samples of size $N/2$ from a normal distribution with a mean of zero and a standard deviation of one and applied the RTI function to this generated data.

From Table 5 one can see that the computational demands for alternation designs using systematic RTI rise exponentially when sample size is increased, allowing for only relatively small sample sizes. However, Table 6 shows that random RTI allows for far larger sample sizes when using alternation designs. Table 5 also shows that the computational demands for phase designs are smaller than for alternation designs, leading to higher usable sample sizes for phase designs when using systematic RTI. Nevertheless, random RTI also allows for larger usable sample sizes when using phase designs compared to systematic RTI.

Conclusion

In this paper, we have illustrated RTI to construct CIs for ES measures in single-case designs without making classic parametric assumptions. Although the ESs we illustrated here were limited to mean differences, RTI can be easily extended to other ES measures and experimental designs. In particular, future research could focus on using RTI to construct nonparametric CIs for nonoverlap statistics in single-case designs.

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

Mieke Heyvaert is a postdoctoral researcher of the Research Foundation – Flanders, (FWO), Belgium (grant ID:1242413N).

Ann Meulders is a postdoctoral researcher of the Research Foundation – Flanders, (FWO), Belgium (grant ID:12E3714N).

References

- Adams, D. C., & Anthony, C. D. (1996). Using randomization techniques to analyse behavioural data. *Animal Behaviour*, *51*, 733–738.
- Allison, D. B., & Gorman, B. S. (1993). Calculating effect sizes for meta-analysis: The case of the single case. *Behaviour Research Therapy*, *31*, 621–631.
- American Psychological Association (1994). *Publication manual of the American Psychological Association* (4th 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: Allyn & Bacon.
- Bauer, D. F. (1972). Constructing confidence sets using rank statistics. *Journal of the American Statistical Association*, *339*, 687–690.
- 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.
- Besag, J., & Diggle, P. J. (1977). Simple Monte Carlo tests for spatial pattern. *Journal of the Royal Statistical Society, Series C*, *26*, 327–333.
- Bulté, I., & Onghena, P. (2008). An R package for single-case randomization tests. *Behavior Research Methods*, *40*, 467–478.
- Bulté, I., & Onghena, P. (2009). Randomization tests for multiple-baseline designs: An extension of the SCRT-R package. *Behavior Research Methods*, *41*, 477–485.
- Bulté, I., & Onghena, P. (2012). When the truth hits you between the eyes: A software tool for the visual analysis of single-case experimental data. *Methodology*, *8*, 104–114.
- Busk, P. L., & Marascuilo, L. A. (1992). Statistical analysis in single-case research: Issues, procedures, and recommendations, with special applications to multiple behaviors. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research designs and analysis: New directions for psychology and education* (pp. 159–185). Hillsdale: Erlbaum.
- Busk, P. L., & Serlin, R. C. (1992). Meta-analysis for single-case research. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 187–212). Hillsdale: Erlbaum.
- Campbell, J. M., & Herzinger, C. V. (2010). Statistics and single subject research methodology. In D. L. Gast (Ed.), *Single subject research methodology in behavioral sciences* (pp. 91–109). New York: Routledge.
- Center, B. A., Skiba, R. J., & Casey, A. (1985–1986). A methodology for the quantitative synthesis of intra-subject design research. *Journal of Special Education*, *19*, 387–400.
- Chambless, D. L., & Ollendick, T. H. (2001). Empirically supported psychological interventions: Controversies and evidence. *Annual Review of Psychology*, *52*, 685–716.
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston: Houghton Mifflin.
- Cox, D. R., & Reid, N. (2000). *The theory of the design of experiments*. Boca Raton: Chapman & Hall/CRC.
- Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.
- DeProspero, A., & Cohen, S. (1979). Inconsistent visual analyses of intrasubject data. *Journal of Applied Behavior Analysis*, *12*, 573–579.

- du Prel, J., Hommel, G., Röhrig, B., & Blettner, M. (2009). Confidence interval or p-value? *Deutsches Ärzteblatt International*, *106*, 335–339.
- Dugard, P. (2014). Randomization tests: A new gold standard? *Journal of Contextual Behavioral Science*, *3*, 65–68.
- Edgington, E. S. (1967). Statistical inference from N=1 experiments. *Journal of Psychology*, *65*, 195–199.
- Edgington, E. S. (1996). Randomized single-subject experimental designs. *Behaviour Research & Therapy*, *34*, 567–574.
- Edgington, E. S., & Onghena, P. (2007). *Randomization tests* (4th ed.). Boca Raton: Chapman & Hall/CRC.
- Emst, M. D. (2004). Permutation methods: A basis for exact inference. *Statistical Science*, *19*, 676–685.
- 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.
- Fisch, G. S. (1998). Visual inspection of data revisited: Do the eyes still have it? *The Behavior Analyst*, *21*, 111–123.
- Garthwaite, P. (2005). Confidence intervals: Nonparametric. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of Statistics in Behavioral Science* (pp. 375–381). Chichester: Wiley.
- Gast, D. L., & Ledford, J. R. (2014). *Single case research methodology: Applications in special education and behavioral sciences* (2nd ed.). New York: Routledge.
- Gibson, G., & Ottenbacher, K. (1988). Characteristics influencing the visual analysis of single-subject data: An empirical analysis. *The Journal of Applied Behavioral Science*, *24*, 298–314.
- Hammond, D., & Gast, D. L. (2010). Descriptive analysis of single-subject research designs: 1983–2007. *Education and Training in Autism and Developmental Disabilities*, *45*, 187–202.
- Harrington, M., & Velicer, W. F. (2015). Comparing visual and statistical analysis in single-case studies using published studies. *Multivariate Behavioral Research*, *50*, 162–183.
- Hartmann, D. P. (1974). Forcing square pegs into round holes: Some comments on “an analysis-of-variance model for the intrasubject replication design”. *Journal of Applied Behavior Analysis*, *7*, 635–638.
- 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.
- Heyvaert, M., & Onghena, P. (2014). Analysis of single-case data: Randomisation tests for measures of effect size. *Neuropsychological Rehabilitation*, *24*, 507–527.
- Heyvaert, M., Moeyaert, M., Verkempynck, P., Van Den Noortgate, W., Vervloet, M., Ugille, & M., Onghena, P. (2016). Testing the intervention effect in single-case experiments: A Monte Carlo simulation study. *Journal of Experimental Education*. doi:10.1080/00220973.2015.1123667
- Hinkelmann, K., & Kempthorne, O. (2008). *Design and analysis of experiments. I and II* (2nd ed.). Hoboken: Wiley.
- 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.
- Hothorn, T., Hornik, K., van de Weil, M. A., & Zeileis, A. (2008). Implementing a class of permutation tests: The coin package. *Journal of Statistical Software*, *28*, 1–23.
- Houle, T. T. (2009). Statistical analyses for single-case experimental designs. In D. H. Barlow, M. K. Nock, & M. Hersen (Eds.), *Single case experimental designs: Strategies for studying behavior change* (3rd ed., pp. 271–305). Boston: Allyn & Bacon.
- Huo, M., & Onghena, P. (2012). RT4Win: A Windows-based program for randomization tests. *Psychologica Belgica*, *52*, 387–406.
- Kazdin, A. E. (2011). *Single-case research designs: Methods for clinical and applied settings* (2nd ed.). New York: Oxford University Press.
- Kendall, M. (1938). A new measure of rank correlation. *Biometrika*, *30*, 81–89.
- Koehler, M. J., & Levin, J. R. (1998). Regulated randomization: A potentially sharper analytical tool for the multiple-baseline design. *Psychological Methods*, *3*, 206–217.
- Kratochwill, T. R., & Levin, J. R. (2010). Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. *Psychological Methods*, *15*, 122–144.
- Kratochwill, T. R., & Levin, J. R. (Eds.). (2014). *Single-case intervention research: Statistical and methodological advances*. Washington, DC: American Psychological Association.
- Kratochwill, T. R., & Stoiber, K. C. (2000). Empirically supported interventions and school psychology: Conceptual and practical issues: Part II. *School Psychology Quarterly*, *15*, 233–253.
- Kratochwill T., Alden, K., Demuth, D., Dawson, D., Panicucci, C., Arntson, P., ... Levin, J. (1974). A further consideration in the application of an analysis-of-variance model for the intrasubject replication design. *Journal of Applied Behavior Analysis*, *7*, 629–633.
- Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). *Single-case designs technical documentation*. Retrieved from What Works Clearinghouse website: http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf
- Kratochwill, T. R., Levin, J. R., Horner, R. H., & Swoboda, C. M. (2014). Visual analysis of single-case intervention research: Conceptual and methodological considerations. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: Methodological and statistical advances* (pp. 91–125). Washington, DC: American Psychological Association.
- LaFleur, B. J., & Greevy, R. A. (2009). Introduction to permutation and resampling-based hypothesis tests. *Journal of Clinical Child & Adolescent Psychology*, *38*, 286–294.
- Lane, J. D., & Gast, D. L. (2014). Visual analysis in single case experimental design studies: Brief review and guidelines. *Neuropsychological Rehabilitation*, *24*, 445–463.
- Lehmann, E. L. (1959). *Testing statistical hypotheses*. Hoboken: Wiley.
- 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.
- 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., Ferron, J. M., & Gafurov, B. S. (2014). Improved randomization tests for a class of single-case intervention designs. *Journal of Modern Applied Statistical Methods*, *13*, 2–52.
- Long, J. D., & Cliff, N. (1997). Confidence intervals for Kendall's tau. *British Journal of Mathematical and Statistical Psychology*, *50*, 31–41.
- Ludbrook, J., & Dudley, H. (1998). Why permutation tests are superior to t and F tests in biomedical research. *The American Statistician*, *52*, 127–132.
- Mann, P. S. (2006). *Introductory statistics*. Hoboken: Wiley.
- Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, *18*, 50–60.
- Manolov, R., & Solanas, A. (2009). Percentage of nonoverlapping corrected data. *Behavior Research Methods*, *41*, 1262–1271.
- Marascuilo, L. A., & Busk, P. L. (1988). Combining statistics for multiple-baseline AB and replicated ABAB designs across subjects. *Behavioral Assessment*, *10*, 1–28.
- 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: Lawrence Erlbaum.

- Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, *105*, 156–166.
- Moore, D. S., McCabe, G. P., & Craig, B. A. (2014). *Introduction to the practice of statistics* (8th ed.). New York: 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 of London. Series A, Mathematical and Physical Sciences*, *767*, 333–380.
- Ongheña, P. (1992). Randomization tests for extensions and variations of ABAB single-case experimental designs: A rejoinder. *Behavioral Assessment*, *14*, 153–171.
- Ongheña, P. (2005). Single-case designs. In B. Everitt & D. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (Vol. 4, pp. 1850–1854). New York: Wiley.
- Ongheña, P., & Edgington, E. S. (1994). Randomization tests for restricted alternating treatments designs. *Behaviour Research and Therapy*, *32*, 783–786.
- Ongheña, P., & Edgington, E. S. (2005). Customization of pain treatments: Single-case design and analysis. *Clinical Journal of Pain*, *21*, 56–68.
- Ongheña, P., & May, R. (1995). Pitfalls in computing and interpreting randomization test p-values: A commentary on Chen & Dunlap. *Behavior Research Methods, Instruments, & Computers*, *27*, 408–411.
- 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., Cryer, J., & Byrns, G. (2006). Controlling trend in single case research. *School Psychology Quarterly*, *21*, 418–440.
- Parker, R. I., Hagan-Burke, S., & Vannest, K. (2007). Percent of all non-overlapping data (PAND): An alternative to PND. *Journal of Special Education*, *40*, 194–204.
- Parker, R. I., Vannest, K. J., & Brown, L. (2009). The improvement rate difference for single case research. *Exceptional Children*, *75*, 135–150.
- Parker, R. I., Vannest, K. J., Davis, J. L., & Sauber, S. B. (2011a). Combining nonoverlap and trend for single-case research: Tau-U. *Behavior Therapy*, *42*, 284–299.
- Parker, R. I., Vannest, K. J., & Davis, J. L. (2011b). Effect size in single-case research: A review of nine nonoverlap techniques. *Behavior Modification*, *35*, 303–322.
- Scruggs, T. E., Mastropieri, M. A., & Casto, G. (1987). The quantitative synthesis of single subject research: Methodology and validation. *Remedial and Special Education*, *8*, 24–33.
- Senchaudhuri, P., Mehta, C. R., & Patel, N. R. (1995). Estimating exact p-values by the method of control variates, or Monte Carlo rescue. *Journal of the American Statistical Association*, *90*, 640–648.
- Shadish, W. R., & Sullivan, K. J. (2011). Characteristics of single-case designs used to assess intervention effects in 2008. *Behavior Research Methods*, *43*, 971–980.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton-Mifflin.
- Sidman, M. (1960). *Tactics of scientific research: Evaluating experimental data in psychology*. New York: Basic Books.
- Solanas, A., Manolov, R., & Ongheña, P. (2010). Estimating slope and level change in N=1 designs. *Behavior Modification*, *34*, 195–218.
- Swaminathan, H., & Rogers, H. J. (2007). Statistical reform in school psychology research: A synthesis. *Psychology in the Schools*, *44*, 543–549.
- Tate, R. L., Perdices, M., Rosenkoetter, U., Wakim, D., Godbee, K., Togher, L., & McDonald, S. (2013). Revision of a method quality rating scale for single-case experimental designs and n-of-1 trials: The 15-item Risk of Bias in N-of-1 Trials (RoBiNT) Scale. *Neuropsychological Rehabilitation*, *23*, 619–638.
- Tritchler, D. (1984). On inverting permutation tests. *Journal of the American Statistical Association*, *385*, 200–207.
- Van den Noortgate, W., & Ongheña, P. (2003). Hierarchical linear models for the quantitative integration of effect sizes in single-case research. *Behavior Research Methods, Instruments, & Computers*, *35*, 1–10.
- Welch, W., & Gutierrez, L. G. (1988). Robust permutation tests for matched-pairs designs. *Journal of the American Statistical Association*, *402*, 450–455.
- Westfall, P. H., & Young, S. S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*. New York: Wiley.
- White, D. M., Rusch, F. R., Kazdin, A. E., & Hartmann, D. P. (1989). Applications of meta-analysis in individual subject research. *Behavioral Assessment*, *11*, 281–296.
- Wilkinson, L., & the APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*, 594–604.
- Ximenes, V. M., Manolov, R., Solanas, A., & Quera, V. (2009). Factors affecting visual inference in single-case designs. *The Spanish Journal of Psychology*, *12*, 823–832.