Behaviour Research and Therapy 117 (2019) 18–27

Contents lists available at ScienceDirect

Behaviour Research and Therapy

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

Randomization tests for changing criterion designs

Patrick Onghena*, René Tanious, Tamal Kumar De, Bart Michiels

KU Leuven, Faculty of Psychology and Educational Sciences, Belgium

ARTICLE INFO

Keywords:
Single-case experiment
Single-case experimental design
N-of-1 trial
Randomization test
Changing criterion design

ABSTRACT

Randomization tests for alternating treatments designs, multiple baseline designs, and withdrawal/reversal designs are well-established. Recent classifications, however, also mention the "changing criterion design" as a fourth important type of single-case experimental design. In this paper, we examine the potential of randomization tests for changing criterion designs. We focus on the rationale of the randomization test, the random assignment procedure, the choice of the test statistic, and the calculation of randomization test *p*-values. Two examples using empirical data and an R computer program to perform the calculations are provided. We discuss the problems associated with conceptualizing the changing criterion design as a variant of the multiple baseline design, the potential of the range-bound changing criterion design, experimental control as an all-or-none phenomenon, the necessity of random assignment for the statistical-conclusion validity of the randomization test, and the use of randomization tests in nonrandomized designs.

Single-case experimental research has a long tradition in the behavioral sciences (Barlow, Nock, & Hersen, 2009; Morley, 2018). In this paper, we focus on one type of design used in single-case experimental research: the changing criterion design. We propose to add random criterion changes to this type of design and discuss the possibility to perform a randomization test on the data collected within such a randomized changing criterion design.

First, we provide a definition of single-case experimental designs and give some historical background. Next, we present the state of the art regarding randomization tests for single-case experimental designs and propose a new randomization test for the changing criterion design. This new randomization test is illustrated with data from two single-case experiments. R programs for all computations can be found in the Appendix.

1. The curious case of single-case experimental research: definition and historical background

A single-case experimental design is a research design in which one case is observed repeatedly during a certain period of time, under different levels of at least one manipulated variable. In psychology, the case is usually a person, but any experimental unit at any level of aggregation can be used. The purpose of using this design is testing the causal effect of one or more manipulated variables on one or more outcome variables for the specific case under investigation (Barlow et al., 2009; Kazdin, 2011; Kratochwill & Levin, 2014b; Ledford & Gast, 2018; Morley, 2018).

The use of single-case experimental designs can be traced back to the origins of psychology as a scientific discipline and to the work of, for example, founding fathers like Wundt, Fechner, Ebbinghaus, Pavlov, Watson, and Skinner (Barlow et al., 2009; Ledford & Gast, 2018; Morley, 2018). The designs were systematically elaborated in the 1960s by researchers working in Skinner's behaviorist tradition (see e.g., Baer, Wolf, & Risley, 1968; Davidson & Costello, 1969; Dukes, 1965; Sidman, 1960), but were not picked up in the general methodological handbooks of that time.

In the 1970s and 1980s excellent and comprehensive overviews were compiled on the topic of both single-case design and analysis (e.g., Hersen & Barlow, 1976; Kazdin, 1982; Kratochwill, 1978), but they were nourished only in the narrow niche of applied behavior analysis. The majority of psychology's methodology was geared towards group comparison designs and inferential statistics, borrowed from agricultural science (see e.g., Kirk, 1995, for an overview). This brought researchers to lament "Where is the individual subject in scientific psychology?" (Valsiner, 1986) or to write opinion articles about single-case designs as "the neglected alternative" (Blampied, 2000) or "the best kept secret" (Lundervold & Belwood, 2000).

2. Turning tables: renewed interest in single-case experimental designs

This picture has changed spectacularly after the first decade in the

* Corresponding author. KU Leuven, Faculty of Psychology and Educational Sciences, Tiensestraat 102, BE-3000, Leuven, Belgium.
E-mail address: patrick.onghena@kuleuven.be (P. Onghena).

https://doi.org/10.1016/j.brat.2019.01.005
Received 15 April 2018; Received in revised form 7 January 2019; Accepted 10 January 2019
Available online 14 January 2019
0005-7967/

21st century. Single-case experimental designs broke loose from their historical ties to behaviorism and found their way in broad university curricula, experimental methodology for a wide diversity of research topics, and general healthcare (Byiers, Reichle, & Symons, 2012; Kratochwill & Levin, 2014b; Lillie et al., 2011; Nikles & Mitchell, 2015; Schork, 2015; Shadish, Kyse, & Rindskopf, 2013; Smith, 2012). Nowadays, single-case experimental designs are thriving in several areas of psychology and one journal after the other is devoting a special issue on the topic (Barker, Mellalieu, McCarthy, Jones, & Moran, 2013; Burns, 2012; Evans, Gast, Perdices, & Manolov, 2014; Howard, Best, & Nickels, 2015; Lenz, 2015; Maggin & Chafouleas, 2013; Maggin, Lane, & Pustejovsky, 2017; Shadish, 2014; Vohra, 2016). What has happened in the meantime to turn the tables?

Interestingly, apart from the continued work on methodology and statistics and a growing number of single-case applications in psychology (see e.g., de Jong et al., 2005; Molenaar, 2004; Onghena & Edgington, 2005; Parker & Brossart, 2003; ter Kuile et al., 2009; Van den Noortgate & Onghena, 2003a, 2003b; Vlaeyen, de Jong, Geilen, Heuts, & van Breukelen, 2001), the major impetus came from methodological developments outside of psychology: from evidence-based medicine and from educational research. In evidence-based medicine, the single-case experimental design has been coined the "*N*-of-1 randomized controlled trial" (*N*-of-1 RCT) and, after considerable scrutiny, these *N*-of-1 RCTs were included among the highest levels of evidence possible to demonstrate therapy effectiveness, both according to the Oxford Centre for Evidence-Based Medicine 2011 Levels of Evidence and in the Evidence-Based Medicine Guidelines of the American Medical Association (Guyatt et al., 2000; Guyatt, Jaeschke, & McGinn, 2002; Howick et al., 2011; Shamseer et al., 2015; Vohra et al., 2015). In educational research, the What Works Clearinghouse (WWC) of the US Department of Education released an influential policy document on Single-Case Design Standards in 2010 that contains well-balanced guiding principles for the design and visual analysis of single-case experiments (Kratochwill et al., 2010, 2013). This document defined the field and its importance for evidence-based education, but also has set the research agenda for the years to come (Kratochwill & Levin, 2014a; 2014b).

Recently, psychology reclaimed the initiative with the Single-Case Reporting guideline In Behavioral interventions (SCRIBE) 2016 Statement (Tate et al., 2016b, 2016a; The University of Sydney, 2018). This Statement was derived using Delphi methodology and comprises the consensus among an impressive panel of 26 international single-case experimental design experts. The simultaneous publication of SCRIBE in 10 journals represents a strong endorsement for the use of single-case experimental designs across the broad field of psychology.[1]

3. Randomization tests for single-case experimental designs

One striking feature of the current guidelines and standards for single-case experimental research is the paucity of clear-cut recommendations for the statistical analysis of single-case data. For example, in a section of the WWC Standards on "Criteria for demonstrating evidence of a relation between an independent variable and an outcome variable", Kratochwill et al. (2010) reported:

> This section assumes that the demonstration of an effect will be established through "visual analysis," as described later. As the field reaches greater consensus about appropriate statistical analyses and quantitative effect-size measures, new standards for effect demonstration will need to be developed. (p. 16)

The lack of a consensus regarding statistical analysis is common in many scientific fields, but in the field of single-case experimental designs it reinforced the sole reliance on so-called "visual analysis" (Horner et al., 2005; Ledford & Gast, 2018; Parsonson & Baer, 1978). In this sense, the SCRIBE 2016 Statement is a major step forward. Tate et al. (2016b) explained: "Both visual and statistical techniques can be used to analyze SCED data (…). They are considered complementary rather than mutually exclusive (…) and should, arguably, be used in combination." (pp. 21–22), and in the Appendix they provide a selection of statistical methods applicable to single-case experimental designs.

In the present article, we want to zoom in on one of the statistical methods mentioned in the Appendix of the SCRIBE 2016 Statement: the use of randomization tests to evaluate the null hypothesis that there is no relation between the manipulated variable(s) and the outcome variable(s) (Tate et al., 2016b, p. 31). The potential of randomization tests in conjunction with single-case experimental designs was first mentioned by Edgington in 1967, based on the permutation test rationale proposed by Fisher (1935), Pitman (1937), and Welch (1937). Randomization tests have the advantage that they do not rely on random sampling or specific population distributions and that they circumvent the problem of serial dependency by conditioning on the observed data (Dugard, File, & Todman, 2011; Edgington & Onghena, 2007; Kratochwill & Levin, 2014b; Onghena, 2018; Park, Marascuilo, & Gaylord-Ross, 1990). Randomization tests derive their statistical validity from the random assignment procedure that was actually carried out while designing the study (Dugard, 2014; Ferron & Onghena, 1996; Ferron & Ware, 1995; Heyvaert & Onghena, 2014b; Houle, 2009; Kazdin, 1984; Morley, 2018). Furthermore, randomization tests can be considered as a generic procedure for constructing a reference distribution so they have the potential to be combined with all kinds of other data-analytical approaches (Ferron & Foster-Johnson, 1998; Ferron & Jones, 2006; Heyvaert et al., 2017; Heyvaert & Onghena, 2014a; Onghena, Michiels, Jamshidi, Moeyaert, & Van den Noortgate, 2018). The main disadvantage of randomization tests is that they are computer-intensive, but with the present-day availability of fast computers, this problem has been largely overcome. The most important remaining obstacle for widespread use of randomization tests for single-case experimental designs is the availability of software for researchers and practitioners (Bulté & Onghena, 2008, 2013; Chen, Peng, & Chen, 2015; De, Michiels, Vlaeyen, & Onghena, 2017; Dugard et al., 2011; Levin, Evmenova, & Gafurov, 2014).

The crucial assumption of randomization tests is that there is an element of random assignment in the study design (Dugard et al., 2011; Edgington & Onghena, 2007; Ferron & Levin, 2014). Of course, random assignment in single-case experimental research cannot refer to the random assignment of participants to treatment groups, as is the case in between-groups randomized controlled trials. Random assignment in single-case experimental research refers to the random assignment of measurement occasions to treatment levels (Edgington, 1996; Heyvaert, Wendt, Van den Noortgate, & Onghena, 2015; Morley, 2018).

The way in which such a random assignment can be accomplished in single-case experimental research differs from design to design (Edgington, 1996; Edgington & Onghena, 2007; Ferron & Levin, 2014). In the SCRIBE 2016 Statement four families of single-case experimental designs have been distinguished: withdrawal/reversal designs, alternating treatments designs, multiple baseline designs, and changing criterion designs (Tate et al., 2016a; 2016b). Based on the specific random assignment procedure and possible test statistics, several

[1] The history of nonexperimental single-case research in developmental and personality psychology parallels this evolution (Danziger, 1990; Hamaker, 2012). When Peter Molenaar published his "Manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever" in 2004, he was overly optimistic. Only few psychologists were paying attention to his manifesto. It took more than another decade before researchers realized what the implications of the so-called "ergodic theorems" were. With these theorems, Molenaar (2004) formally showed that what applies in aggregate (by taking averages) is not necessarily informative for what is true for individuals in general. The consequences of these theorems for human subjects research are far-reaching (Fisher, Medaglia, & Jeronimus, 2018; Onghena et al., 2018).

randomization tests have already been proposed for withdrawal/reversal designs, alternating treatments designs, and multiple baseline designs (see e.g., Bulté & Onghena, 2009; Edgington, 1975, 1980b; Levin, Ferron, & Gafurov, 2016, 2017; Manolov & Onghena, 2018; Onghena, 1992; Onghena & Edgington, 1994). By contrast, randomization tests for changing criterion designs have never been given due consideration, although the SCRIBE 2016 Statement explicitly mentions that "each of the single-case experimental designs has the capacity to introduce randomization into the design" (Tate et al., 2016a, p. 4) and although *randomized* changing criterion designs have their proper place in the SCRIBE design classification (Tate et al., 2016a; 2016b).

4. Randy

Before we propose our new randomization test for the changing criterion design, it might be instructive to recapitulate how the "old" randomization test in a more common single-case experimental design works. In this section, we follow the basic rationale of the single-case randomization test as proposed by Edgington in 1967.

Suppose we want to test side effects of sleeping medication in a 53-year-old male, called "Randy". Randy takes sleeping pills every evening, but during the last weeks he is experiencing severe dizziness when he wakes up in the morning. He contacts his physician with this complaint and together they wonder whether the dizziness might be related to the use of the sleep medication. In order to investigate whether there are indications that this relation holds, they agree to conduct an experiment with Randy as the sole participant. They agree to perform a 6-day experiment, alternating 3 days of sleeping medication and 3 days of placebo. Randy will record his feelings of dizziness on a 7-point Likert scale, going from 1 = no dizziness, to 7 = very severe dizziness.

Of course, they will not alternate the 3 days of sleeping medication and the 3 days of placebo in a systematic way that both know beforehand. Such a predetermined unblinded method could cause all sorts of experimenter and expectancy confounding effects. Instead, they consult with a pharmacist who prepares the placebo and active medication capsules, randomizes their order, and keeps the randomization schedule to himself. This makes a randomized double-blind placebo-controlled single-case experiment possible.

After one week, Randy returns to the physician with his 6-day record. His scores for the consecutive days are 5, 6, 7, 4, 1, and 6. How can we evaluate these scores? First, the pharmacist is called to reveal the randomization schedule. This schedule turns out to be A B B A A B, with A = Placebo and B = Sleeping medication. Second, a summary statistic is computed. For example, the difference between the average dizziness scores for the two treatments is equal to

$$\bar{B} - \bar{A} = \frac{19}{3} - \frac{10}{3} = 3$$

This means that, on average, the dizziness score on the active medication days is 3 points higher than on the placebo days. Third, because a randomized experiment has been carried out, it is possible to calculate a randomization test *p*-value. This *p*-value is calculated under the assumption that there is no relation between the treatment and the scores (the null hypothesis). If this assumption is true, then the six scores would have been obtained no matter which treatment was given on that particular day. This means that, given the true null hypothesis and given the randomization, the scores are fixed and the treatments labels (A and B) are random in virtual replications of the experiment. We can see that there are 6!/3!3! = 20 ways to randomly order three A's and three B's, and therefore there are 20 ways to carry out the experiment (of which the observed experiment is only one instance).

Suppose we now calculate the summary statistic $\bar{B} - \bar{A}$ for each of the 19 other possible randomizations and that we locate the observed value of 3 for the actual experiment in the distribution of values for all possible randomizations. Table 1 shows the result of these calculations. The observed value of 3 for the actual experiment is the largest value in the set of 20 values. In probability language, this is saying that the probability to obtain a difference between means of 3 or larger is equal to 0.05 ($p = .05$) if the null hypothesis were true. Randy and the physician can take this *p*-value into account when taking a decision regarding the causal link between the sleeping pills and the dizziness and ultimately when taking a decision regarding stopping sleeping medication or changing treatment. They could use Fisher's disjunctive argument that an unlikely event has occurred OR that the null hypothesis is false. If they would have agreed to rule out events with a probability equal to or smaller than 5% (the significance level), then they could formally reject the null hypothesis.[2]

Table 1
All possible randomization schedules and values of the test statistic $\bar{B} - \bar{A}$ for a single-case experimental design with two treatment conditions, three measurement occasions each, and the observed scores 5, 6, 7, 4, 1, 6.

Design	$\bar{B} - \bar{A}$	Design	$\bar{B} - \bar{A}$
A A A B B B	−2.33	B B B A A A	2.33
A A B A B B	−0.33	B B A B A A	0.33
A A B B A B	1.66	B B A A B A	−1.66
A A B B B A	−1.66	B B A A A B	1.66
A B A A B B	−1.00	B A B B A A	1.00
A B A B A B	1.00	B A B A B A	−1.00
A B A B B A	−2.33	B A B A A B	2.33
*A B B A A B	*3.00	B A A B B A	−3.00
A B B A B A	−0.33	B A A B A B	0.33
A B B B A A	1.66	B A A A B B	−1.66

Note. The design actually used in the example and the observed value of the test statistic are indicated with an asterisk.

5. Randomization tests for changing criterion designs

The example above capitalizes on the ability to quickly alternate treatment conditions, to have a rapid onset of the effect, no need for wash-out, and double-blind treatment allocation. This is usually difficult, undesirable, or impossible for behavioral interventions. For behavioral interventions, we need longer phases to implement the treatment, effects may be gradual and permanent, and the treatment cannot (or should not) be concealed. For all these reasons, alternative single-case experimental designs without fast alternations have been proposed. The changing criterion design is one of them.

In their seminal paper, Hartmann and Hall (1976) introduced the changing criterion design as a single-case experimental design in which a person tries to reach a goal (or "criterion") in several consecutive discrete steps. A changing criterion design consists of an initial baseline period, followed by a series of treatment phases that are associated with stepwise (increasing or decreasing) changes in the criterion. The causal effect in the changing criterion design is inferred from the similarity between the criteria (as levels of the manipulated variable) and the levels reached in the outcome variable.

Hartmann and Hall (1976) illustrated the changing criterion design with two single-case experiments from their own therapeutic work. We will use part of the data of the first experiment to demonstrate how a randomization test can be constructed for a changing criterion design. In this experiment, a behaviorally disordered boy was asked to solve mathematical problems. In the baseline phase (Phase A), the number of correct solutions was registered during several consecutive math sessions, without setting any criterion. After the baseline phase, the criterion was set at two correct solutions (Phase B), with the consequence

[2] The example is kept simple for illustrative purposes. Readers who want more extensive background on the randomization test rationale are referred to Edgington and Onghena (2007) and Onghena (2018). Issues regarding the design and analysis of alternating treatment designs are discussed in Onghena and Edgington (1994) and Manolov and Onghena (2018).

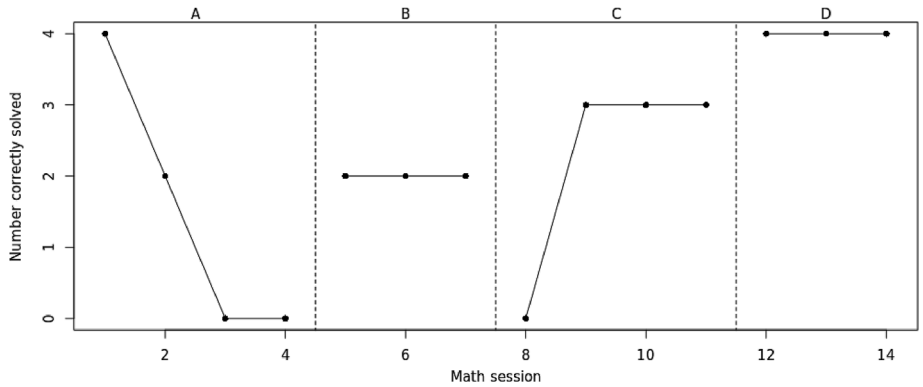

Fig. 1. Number of correctly solved mathematical problems by a behaviorally disordered boy during 14 math sessions, divided in an A phase (no reward), a B phase (reward at two correct solutions), a C phase (reward at three correct solutions), and a D phase (reward at four correct solutions).

that reaching this criterion gave access to recess and the opportunity to play basketball. In the second treatment phase (Phase C), an identical reward was given but with a criterion of three correct solutions. And finally in the third treatment phase (Phase D), the reward was only given after the criterion of four correct solutions was met. Each phase consisted of three or four math sessions. The results for the four phases can be found in Fig. 1.

The purpose of using this design is to infer a causal effect between the criteria for the reward and the number of correct solutions, or in the words of Hartmann and Hall (1976): "When the rate of the target behavior changes with each stepwise change in the criterion, therapeutic change is replicated and experimental control is demonstrated" (p. 527).

As in the Randy example, the two main ingredients of a randomization test to infer such a causal effect from these data are (1) a test statistic, and (2) a random assignment procedure. Based on these ingredients, (3) a randomization test *p*-value can be derived.

5.1. The test statistic

Randomization tests are flexible with respect to the test statistics that can be used (Edgington & Onghena, 2007; Heyvaert & Onghena, 2014a). This means that any summary statistic or effect size measure is a potential test statistic, given that it is sensitive to the kind of effect that is expected. In the Randy example, the difference between means was used to express the size of the effect. For the changing criterion design, which by definition has many treatment levels, the choice of the test statistic is not that obvious. Taking into account the logic of the changing criterion design, we consider three potentially interesting effect size measures that also have the added advantage of intuitive appeal: the Pearson product-moment correlation coefficient, the Proportion of Conforming Data, and the Mean Absolute Deviation.

- The Pearson product-moment correlation coefficient between the outcome scores and the criterion levels could be taken as a measure of effect size in a changing criterion design. This effect size measure is implicit in articles and books presenting the changing criterion design with an emphasis on the assessment of "covariation" between scores and criteria (see e.g., Kratochwill et al., 2010, p. 6). In the example of Fig. 1, the criteria are 2, 3, and 4 correct solutions for the B, C, and D phase respectively. If we would take the median score of 1 in the baseline phase as the first criterion, then this measure of effect size yields a value of 0.5698 for the data in Fig. 1. This value would be considered large, using Cohen's (1988, 1992) rules of thumb.
  A major drawback of the Pearson product-moment correlation coefficient, however, is that scores that are linearly increasing (or decreasing) alongside the criteria will result in large values, but that the level of the scores is not taken into account. In other words, no perfect match is needed between the scores and the criteria, while the strength of the changing criterion design for making causal inferences relies exactly on this match. Consequently, the Pearson product-moment correlation coefficient gives an indication of the linear association between the scores and the criteria, but is not very sensitive to the kind of effect that is expected in changing criterion designs. Therefore, we do not recommend the Pearson product-moment correlation coefficient as a test statistic for routine use in a changing criterion design.
- Another effect size measure that has been proposed for the changing criterion design is the Proportion of Conforming Data (PCD, see e.g., McDougall, 2005, 2013). PCD is calculated by counting the number of scores that are equal to the criterion, divided by the total number of scores. In the running example with 1, 2, 3, and 4 as the criteria, this measure of effect size yields a value of $9/14 = 0.6429$.
  This measure is more in line with the logic of the changing criterion design than the Pearson product-moment correlation coefficient, but it does not take the magnitude of the difference between the scores and the criterion into account. Therefore, this measure is also not very sensitive as a test statistic. We would recommend this measure only if the magnitude of the difference between the scores and the criterion is irrelevant or if a test is needed that is robust against outliers, at the expense of general sensitivity.
- A measure of effect size that shares PCD's matching logic, but that is also sensitive to the magnitude of the difference between the scores and the criterion is the Mean Absolute Deviation (MAD) between the scores and the criteria. The MAD is calculated by taking the difference between the scores and the criterion at each measurement occasion, dropping the sign, summing all these absolute differences, and dividing by the total number of scores. In the running example, $MAD = 3 + 1 + 1 + 1 + 0 + 0 + 0 + 3 + 0 + 0 + 0 + 0 + 0 + 0 = 9/14 = 0.6429$.
  Because the MAD is a deviance score, smaller values represent a better fit between the scores and the criteria, with 0 as the perfect match. The measure has no upper bound, but because the outcome variables in changing criterion designs are, by definition, manifest and countable, the interpretation of MAD should be straightforward. For example, the observed value of 0.6429 for the data in Fig. 1 means that the number of correct solutions deviates less than 1 solution from the criterion, on average, which could be a satisfactory result, depending on the boy's disorder, the difficulty of the math problems, and the availability or importance of the reward.

One issue with defining an overall effect size measure for the changing criterion design is the inclusion of the baseline phase in the calculations. In the example above, we used the median of the scores in the baseline phase as the criterion, but other options are possible. One

could argue that there is no criterion in the baseline phase, and hence the criterion should be 0, but this is evidently not sensible in most applications that have a general baseline level far above 0. Furthermore the criteria represent benchmarks around which the scores are expected to vary; 0 is the minimum value in the baseline phase and cannot be considered such a benchmark. Another option is to leave the baseline scores out of all calculations, but this wastes important information because the phase change from the baseline to the first treatment surely is indicative of the experimental control exerted by the criterion. The option for using the median of the scores in the baseline phase as the criterion seems to be a good compromise and incorporates stability of the baseline in the effect size measure (or the instability in the example). More stable baselines result in more confidence about the causal effect between the manipulated variable and the outcome variable (Barlow et al., 2009; Kazdin, 2011; Klein, Houlihan, Vincent, & Panahon, 2017), and this should be reflected in the effect size measure and the statistical test.

5.2. The random assignment procedure

We are used to thinking about random assignment in terms of randomly assigning participants to treatment levels. However, this scheme does not work in a single-case experimental design because there is only one participant. Instead, the random assignment procedure in a single-case experimental design refers to the random assignment of *measurement occasions* to treatment levels (Edgington, 1996), as was illustrated in the Randy example. In a changing criterion design, there is an additional complication that the random assignment procedure cannot refer to *unrestricted* random assignment of *all* measurement occasions to *all* treatment levels. The changing criterion design has its own specific design structure with consecutive phases and criteria, and this structure needs to be preserved if we want to evaluate the observed result against a set of results that would be obtained if another schedule for the actual experiment was randomly selected, just as we did in the Randy example. To make this more concrete: in the study of Fig. 1 we have four A observations, three B observations, four C observations, and three D observations. We cannot just randomly pick one design by shuffling the 14 phase indicators and select one of the 14!/4!3!4!3! = 4,204,200 combinations. A randomly picked design, such as CBCAADABBCDDAC would make it impossible to conduct the experiment as a changing criterion design.

A random assignment procedure that preserves the design structure of the changing criterion design involves the random determination of the phase change points in the series of measurement occasions. This procedure was originally proposed by Edgington (1975) for simple AB designs, but can be extended to changing criterion designs (for other extensions, see also Levin, Ferron, & Gafurov, 2014; Onghena, 1992; Onghena, Vlaeyen, & de Jong, 2007). In general, the random assignment procedure for the changing criterion design involves defining a population of potential phase change points and randomly selecting the phase change points for the actual experiment. For the experiment of Fig. 1, suppose that the population of potential phase change points consisted of measurement occasions 4 or 5 for the B phase, measurement occasions 7 or 8 for the C phase, and measurement occasions 11 or 12 for the D phase. For the actual experiment, we have to randomly select one of the measurement occasions for each phase change. There are $2^3 = 8$ possibilities to randomly select one of the two measurement occasions for each of the three phase change points. In this way, the random assignment is restricted to eight possibilities instead of the 4,204,200 combinations for unrestricted randomization, but all experiments would be feasible and can follow the changing criterion design logic. Suppose that, in the example, the (randomly) selected start points are 5, 8, and 12. The other possibilities are listed in the two first columns of Table 2.

Actually, a randomized version of the changing criterion design might be more in line with the logic of this design than a nonrandomized version. Klein et al. (2017) identified the use of phases with variable length as best practice in changing criterion designs. Fixed phase lengths carry the risk of covariation with other cyclical time trends (e.g., diurnal, weekly, monthly, or seasonal cycles). Variable phase lengths can demonstrate a causal link between the criterion and the outcome variable, independent of the phase length, the number of repeated measurements, and confounding variables associated with these features. Rather than systematically varying the time lengths, it would be a small effort for the researcher to list all feasible starting points and to randomly pick one.

In some applications, it might be recommended to wait for the start of the B phase until the baseline is stable or to wait for a change of phase after the previous criterion has been met at one or more measurement occasions. As Edgington (1980a) has demonstrated, the random assignment procedure can be modified to accommodate such response-guided design prescriptions. In such cases, the random determination of the start points is postponed until the baseline is stable or the criterion has been met. Afterwards, a valid randomization test is still possible, following the steps described in the next sections. In any case, the random assignment procedure (restricted and possibly postponed) is important because it guides the construction of the reference distribution, as will be explained in the next sections.

Table 2
Possible start points for the B, C, and D phase in a randomized version of the changing criterion design in Fig. 1, the resulting design with phase labels for each measurement occasion, and the corresponding Pearson product-moment correlation coefficient (COR), proportion of conforming data (PCD), and mean absolute deviation (MAD).

Start points	Design	*COR*	*PCD*	*MAD*
(4, 7, 11)	A A A B B B C C C C D D D D	0.4810	0.5000	0.8571
(4, 7, 12)	A A A B B B C C C C D D D D	0.4777	0.5714	0.7857
(4, 8, 11)	A A A B B B B C C C D D D D	0.4938	0.5714	0.7857
(4, 8, 12)	A A A B B B B C C C C D D D	0.4934	0.6429	0.7143
(5, 7, 11)	A A A A B B C C C C D D D D	0.5517	0.5000	0.7857
(5, 7, 12)	A A A A B B C C C C C D D D	0.5528	0.5714	0.7143
(5, 8, 11)	A A A A B B B C C C D D D D	0.5656	0.5714	0.7143
(5, 8, 12)*	A A A A B B B C C C C D D D*	0.5698*	0.6429*	0.6429*

Note. The design actually used in Fig. 1 and the observed test statistics are indicated with an asterisk.

5.3. The randomization test p-value

We stated above that the null hypothesis of the randomization test is that there is no relation between the manipulated variable(s) and the outcome variable(s). For the changing criterion design, this null hypothesis can be reformulated as saying that there is no causal effect of the level of the outcome specified as a criterion on the observed level of the outcome. If this null hypothesis is true, then it does not matter which start points are selected: The outcome scores would always be the same. In other words, if the null hypothesis of no causal effect is true, then the outcome scores are fixed values for each measurement occasion, no matter which phase they are in.

Therefore, given a true null hypothesis, we can calculate the probability to obtain a value of the test statistic that is as extreme as, or more extreme, than the observed value of the test statistic. The reasoning is identical as in the Randy example. For the calculation of the test statistic values for each of the random assignment possibilities, we need to know the outcome scores and we need to know the start points. As mentioned above, again given a true null hypothesis, the outcome scores are fixed values for each measurement occasion. Given the random assignment, the start points are random. The probability to pick a particular set of start points is 1/8 (in general: 1 divided by the number of random assignment possibilities) and therefore the probability for the associated value of the test statistic is also 1/8. All potential values of the test statistic form a reference distribution and the

observed test statistic can be located within this reference distribution. The randomization test *p*-value is the tail probability of this reference distribution, and is calculated as the sum of the probabilities associated with values of the test statistic that are as large as, or larger than (for the right tail), or as small as, or smaller than (for the left tail), the observed test statistic.

The choice of the tail depends on the test statistic. If large values are indicative of a causal effect, then the right-tail *p*-value is relevant. If small values are indicative of a causal effect, then the left-tail *p*-value is relevant. In general, also two-tailed *p*-values can be defined for randomization tests, but in the context of a changing criterion design two-tailed *p*-values will rarely be useful because the design logic is strongly based on directional predictions and exact matching between the scores and the criteria.

In the example, the Pearson product-moment correlation coefficient and the PCD need a right-tail *p*-value. The MAD needs a left-tail *p*-value. Table 2 shows the test statistics that would have been obtained for each of the start points. For the Pearson product-moment correlation coefficient, the observed statistic is the largest, hence the *p*-value is $1/8 = 0.125$. For the PCD, there is one other test statistic that is equal to 0.6429, hence the *p*-value is $2/8 = 0.250$. The observed test statistic for the MAD is the smallest, hence the *p*-value is $1/8 = 0.125$.

Note that in the given example, the number of random assignment possibilities is kept very small to perform the calculations by hand and to enable enumeration of all test statistics in a small table. This has as a consequence that the statistical power to detect a causal effect between the criterion and the scores is zero for any traditional significance level because the lower bound for the *p*-value is .125. In real-life applications of the changing criterion design, though, the number of phases and the number of measurement occasions in each phase is usually much larger. If the number of measurement occasions in each phase is larger, then the number of potential start points can be increased without compromising the minimum number of measurement occasions in each phase. For example, if the study of Fig. 1 and Table 2 is extended with an E and F phase, and if there are four potential start points for each phase, then the number of random assignment possibilities is already $4^5 = 1024$, making the minimal *p*-value smaller than 0.001.

6. A more elaborate example

The second experiment reported by Hartmann and Hall (1976) is more elaborate. In this experiment a changing criterion design is applied for a heavy smoker in a smoking reduction program. The treatment consisted of imposing a decreasing criterion for the number of cigarettes that the person was allowed to smoke on a day-to-day basis. In a one-week baseline phase, the number of cigarettes was merely registered. After that baseline phase, a stepwise reduction was aimed at: 46 cigarettes in Phase B, 43 in Phase C, 40 in Phase D, 38 in Phase E, 36 in Phase F, and 34 in Phase G. If the participant smoked more cigarettes than the criterion, then a $1.00 fine for each cigarette had to be paid. If the participant smoked fewer cigarettes than the criterion, then a $0.10 bonus for each cigarette was received.

The results of the experiment can be found in Fig. 2. The data were recovered from Fig. 2 of Hartmann and Hall (1976) using "GetData Graph Digitizer" version 2.26 (Fedorov, 2013). Hartmann and Hall (1976) indicate that this person ran through 24 phases in the actual smoking reduction program but in Hartmann and Hall (1976) no data are reported on the remaining phases.

A randomization test on these data is possible if we assume that some form of random assignment of the phase change points has taken place. Suppose that the phase change points could have been any of the three days before or after the actually observed phase change points, making a total of seven potential phase change points for each of the six changes. We assume that the design that was actually used for the study is just a randomly selected design out of the $7^6 = 117{,}649$ possibilities. Suppose that we use the MAD as the test statistic and that we use the median as the criterion for the baseline phase (49 cigarettes in this case) because we want to include the baseline information in our calculations. Suppose we set the level of significance at 5%.

A randomization test on these data is not possible by hand because all 117,649 possibilities have to be evaluated. A computer program in R that performs these calculations is given in the Appendix. Note that even on a fast present-day personal computer this program takes a few minutes to run because of the large computational burden. Valid (but less precise) randomization tests can be carried out by evaluating only a random subsample of all possibilities to have faster output (so-called "Monte Carlo randomization tests", see Edgington & Onghena, 2007), but in changing criterion designs the number of possibilities is usually within the computational reach of present-day computers (given a little bit of patience on the part of the user), and so "exhaustive" randomization tests are recommended.

The observed MAD is 0.8132, which means that the number of cigarettes smoked daily deviates less than 1 cigarette from the criterion, on average. This seems to be a convincing result, given that we are dealing with a heavy smoker (and associated health risks), the criteria dropped from 46 to 34, and the monetary penalty is small. In addition, this observed value can be compared to the other 117,648 values that would have been obtained if other start points had been selected and if the null hypothesis of no treatment effect were true. As can be seen in Fig. 3, the observed MAD value of 0.8132 is located in the far left tail of the reference distribution. The computer program indicates that there are 973 MAD values that are smaller than the observed value, which gives a randomization test *p*-value of $973/117649 = 0.0083$. We can conclude that there is a statistically significant match between the criteria and the number of cigarettes smoked, which is indicative of experimental control and corroborates a causal link between the treatment and the outcome for this heavy smoker. For illustrative purposes, we can also add that the PCD for these data is 0.7033, with a randomization test *p*-value of $246/117649 = 0.0021$ and that the Pearson product-moment correlation coefficient for these data is 0.9388, with a randomization test *p*-value of $14752/117649 = 0.1254$.

7. Discussion

The aim of the present article was to apply the randomization test rationale to one type of single-case experimental design that was identified in recent classifications as having distinct features: the changing criterion design. We used two data sets from the seminal article by Hartmann and Hall (1976) to illustrate the approach and the computations.

We focused on data-analytical issues because of this focus on randomization tests but many other methodological issues are important in evaluating a design and the resulting causal inference. The reader is referred to the excellent review of Klein et al. (2017) for a discussion of these issues in changing criterion designs. Furthermore, it should be acknowledged that in real-life applications additional qualitative details about the procedure, the participant, and the context are important for gauging the credibility and transferability of the results (Onghena, Maes, & Heyvaert, in press).

Sometimes practical, ethical or feasibility issues of this real-life context put restraints on the data-analytical techniques that can be applied. In this sense, randomization tests are flexible data-analytical tools that can be modified for a diversity of designs and test statistics, but they are not a panacea. We want to emphasize that our focus on randomization tests in this article does not represent an idée fixe on *p*-values. We regard randomization test *p*-values as information that is complementary to the results of a visual analysis, the use of effect size measures, and additional qualitative data. Moreover, the randomization test rationale can also be used to construct confidence intervals for effect size measures, because a confidence interval can be considered an interval of values for which the null hypothesis cannot be rejected. Repeatedly performing a randomization test for an increasing or

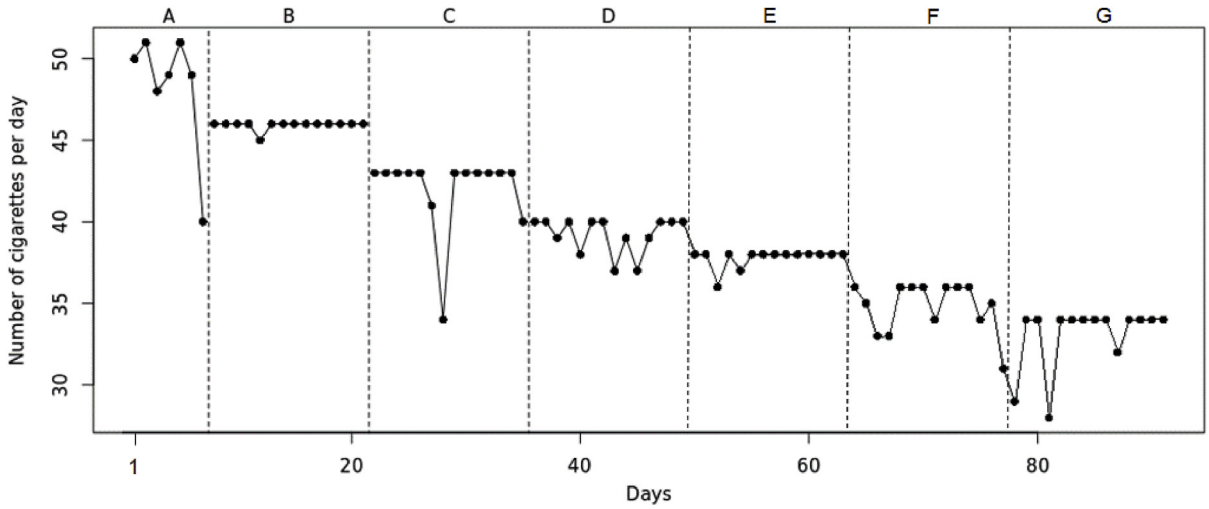

Fig. 2. Number of cigarettes smoked during 91 days in the second experiment of Hartmann and Hall (1976) using a changing criterion design, divided in an A phase (baseline), a B phase (criterion set at 46 cigarettes), a C Phase (criterion set at 43 cigarettes), a D Phase (criterion set at 40 cigarettes), an E Phase (criterion set at 38 cigarettes), an F Phase (criterion set at 36 cigarettes), and a G Phase (criterion set at 34 cigarettes).

decreasing fine-grained series of effect size values can demarcate such an interval (for an elaboration of this idea and accompanying software, see Michiels, Heyvaert, Meulders, & Onghena, 2017).

In the remainder of this article we want to discuss five issues related to the randomization test rationale that we proposed: (1) the changing criterion design as a variant of the multiple baseline design, (2) the range-bound changing criterion design proposed by McDougall (2005), (3) experimental control as an all-or-none phenomenon, (4) the necessity of random assignment for the statistical-conclusion validity of the randomization test, and (5) the use of randomization tests in non-randomized designs.

1. Hartmann and Hall (1976) conceptualized the changing criterion design as a variant of the multiple baseline design. In their analysis, they proposed to use all data before each phase change as baseline data for comparing all data after each phase change. For example, in Fig. 2 this would imply that we can analyze these data as a multiple baseline design with six tiers. If we follow this line of reasoning, we could merely apply the existing multiple baseline design randomization tests to this variant. However, this conception of the changing criterion design as a variant of the multiple baseline design has been debunked (see e.g., Cooper, Heron, & Heward, 2007; Kazdin, 2011; Klein et al., 2017). As Klein et al. (2017) put it: “the behavior being

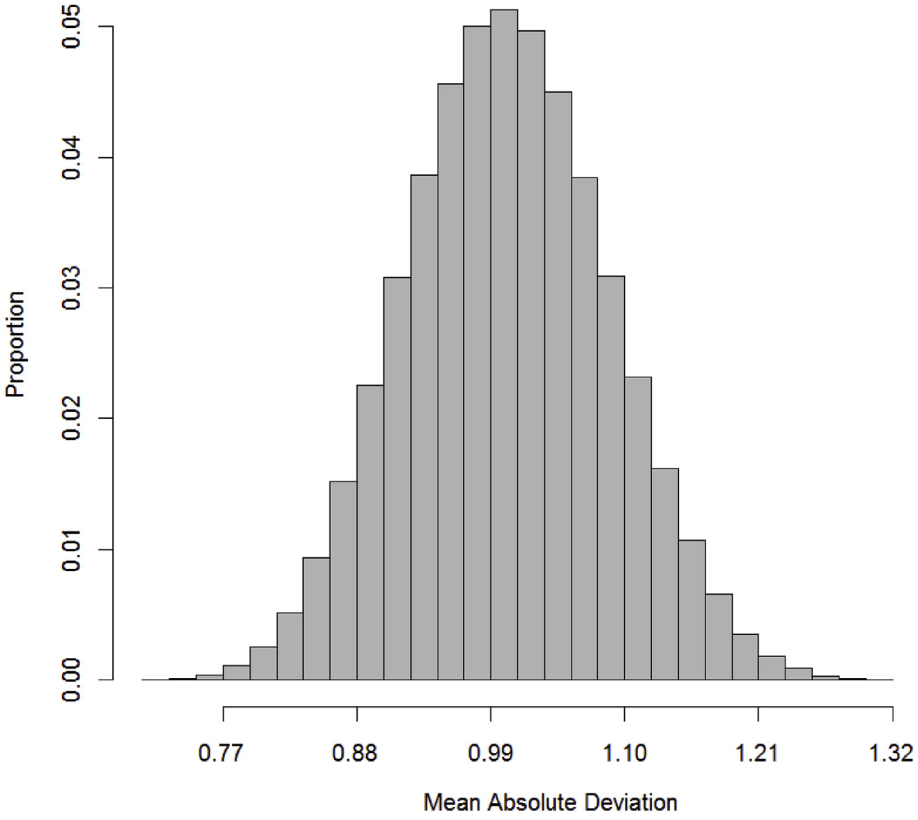

Fig. 3. Reference distribution for the Mean Absolute Deviation statistic using a randomization test on the data of Fig. 2 with randomly shifting start points.

evaluated in a CCD [changing criterion design] is intentionally interdependent (unlike a multiple baseline design) which can lead to an exaggerated perception of experimental control. Conceptualizing this as an extension of the multiple baseline design adds confusion" (p. 52). Another way to make the same point: Analyzing a changing criterion design as a multiple baseline design boils down to using the same data multiple times (e.g., six times the same data in Fig. 2). If a test would be applied without adjusting for this multiple use, then much smaller *p*-values will be obtained than justified by the design and the data.

2. An appealing variant of the changing criterion design was proposed by McDougall (2005) and the randomization test that we proposed can also be applied to this variant. In this variant, the researcher sets a lower bound and an upper bound to the criterion in each phase instead of setting a single fixed point. The aim is to keep the outcome scores within the range, defined by the lower and upper bound. The benefits of using such a "range-bound changing criterion design" include avoiding counter-therapeutic consequences of excessive improvements (e.g., in sports and training) and making the interpretation of the results more straightforward by being explicit about the range of acceptable outcome scores (McDougall, 2013). The only modification for the randomization test that we proposed in order to be applicable for range-bound changing criterion designs refers to the test statistic. For the MAD, all values within the range are coded as having a deviation of 0 and all values outside the range are coded as the absolute difference with the upper or the lower bound, respectively. For the PCD, there is no modification needed: all values within the range are coded as 1 and all values outside the range are coded 0.
3. Single-case researchers, who eyeball graphs as the dominant technique for data analysis, have a tendency to consider experimental control as an all-or-none phenomenon. The result of visual analysis is usually a demonstration of experimental control in absolute terms (see e.g., Barlow et al., 2009; Hartmann & Hall, 1976; Kazdin, 2011; Klein et al., 2017; McDougall, 2005, 2013). By contrast, experimental control in real-life applications is more gradual and outcome scores show much more variability than the textbook examples. Effect size measures and randomization tests can reflect this gradualness. A well-chosen effect size measure gives an indication of the strength of the association between the criterion and the outcome scores and can be used as a test statistic in a randomization test (Heyvaert & Onghena, 2014a). Also randomization tests do not result in absolute and deterministic conclusions, but result in a tentative and probabilistic causal inference based on counterfactual reasoning (Onghena, 2018). The growing consensus that the best practice is to combine visual exploration, quantitative description, and statistical inference (Bulté & Onghena, 2012; Manolov, Gast, Perdices, & Evans, 2014; Manolov & Moeyaert, 2017; Morley, 2018; Tate et al., 2016a, 2016b) will eventually also mitigate dichotomous thinking with respect to experimental control.
4. The necessity to implement some form of random assignment procedure for the valid use of randomization tests is frequently mentioned as the major drawback for the broad applicability of these tests. Single-case researchers do not regularly use random assignment procedures and may be reluctant to do this in the future because these procedures are perceived as conflicting with the deterministic experimental control logic (see e.g., Barlow et al., 2009; Kazdin, 1980; Ledford & Gast, 2018). We want to argue that, although random assignment requires an additional effort in planning the experiment, the inclusion of random assignment first and foremost strengthens the internal validity of the design (Dugard et al., 2011; Edgington, 1996; Heyvaert et al., 2015; Kratochwill & Levin, 2010; Tate et al., 2013). In this regard, random assignment in single-case experimental designs is as important (or controversial) as random assignment in between-groups randomized controlled trials (Onghena, 2018). For sure, *N*-of-1 RCTs would never have been included among the highest levels of evidence in evidence-based medicine without the random assignment component (Guyatt et al., 2000, 2002; Howick et al., 2011; Shamseer et al., 2015; Vohra et al., 2015). Furthermore, randomly selecting the start points of the phases might be very natural in the changing criteria design. As we mentioned above, the use of phases with variable length is considered best practice for changing criterion designs (Klein et al., 2017). One obvious and undemanding technique for arriving at changing criterion designs with variable phase lengths is to randomly select phase start points from a limited set of potential phase start points.
5. If there are insurmountable obstacles to using some form of random assignment, then we believe that there is still a place for a descriptive use of randomization tests. This "descriptive use" refers to the fact that there is no stochastic basis for any inference. The randomization test *p*-value in a nonrandomized design becomes just another quantitative description of the data, alongside other descriptive statistics, such as effect size measures. These *p*-values might represent a reasonable addition to these effect size measures because a *p*-value is a familiar metric with a known range. The calculation of randomization test *p*-values in nonrandomized designs can be considered as performing a monotonous transformation of an effect size measure to the 0–1 range. If the effect size measure can be interpreted, then there is also a meaningful interpretation for the *p*-value. For example, the data in Fig. 2 were collected with a nonrandomized design but nevertheless the visual impression of the treatment effect is similar to mentally shifting the phase changes a number of days to the left and a number of days to the right and noticing that the effect is strongly associated with the change points that were actually used. In the interpretation of such a "nonrandom" *p*-value we have to be careful that our reach does not exceed our grasp, though. It is not because a *p*-value is calculated, that a valid statement about statistical significance can be made. In this respect, the use of randomization tests in nonrandomized designs is similar to the use of parametric tests in the absence of random sampling (Winch & Campbell, 1969).

8. Conclusion

Randomization tests represent a versatile technique in the toolbox of the single-case researcher. We showed in this article how they can be applied in changing criterion designs and we explained how they can be modified for range-bound changing criterion designs. Randomization tests have become practical with the wider availability of fast personal computers and user-friendly software, and in the Appendix we provided an R computer program for randomization tests in changing criterion designs and range-bound changing criterion designs. We recommend to include a random assignment procedure when determining the phase changes in the design and to use the R computer program with caution in nonrandomized designs.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.brat.2019.01.005.

References

Baer, D. M., Wolf, M. M., & Risley, T. R. (1968). Some current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis, 1*, 91–97. https://doi.org/10.1901/jaba.1968.1-91.

Barker, J. B., Mellalieu, S. D., McCarthy, P. J., Jones, M. V., & Moran, A. (2013). Special Issue on single-case research in sport psychology. *Journal of Applied Sport Psychology, 25*, 1–3. https://doi.org/10.1080/10413200.2012.729378.

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

Blampied, N. M. (2000). Single-case research designs: A neglected alternative. *American Psychologist, 55*, 960. https://doi.org/10.1037/0003-066X.55.8.960.

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

Bulté, I., & Onghena, P. (2009). Randomization tests for multiple baseline designs: An extension of the SCRT-R package. *Behavior Research Methods, 41*, 477–485. https://doi.org/10.3758/BRM.41.2.477.

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. https://doi.org/10.1027/1614-2241/a000042.

Bulté, I., & Onghena, P. (2013). The Single-Case Data Analysis package: Analysing single-case experiments with R software. *Journal of Modern Applied Statistical Methods, 12*, 450–478. https://doi.org/10.22237/jmasm/1383280020.

Burns, M. K. (2012). Meta-analysis of single-case design research: Introduction to the special issue. *Journal of Behavioral Education, 21*, 175–184. https://doi.org/10.1007/s10864-012-9158-9.

Byiers, B. J., Reichle, J., & Symons, F. J. (2012). Single-subject experimental design for evidence-based practice. *American Journal of Speech-Language Pathology, 21*, 397–414. https://doi.org/10.1044/1058-0360.

Chen, L.-T., Peng, C.-Y. J., & Chen, M.-E. (2015). Computing tools for implementing standards for single-case designs. *Behavior Modification, 39*, 835–869. https://doi.org/10.1177/0145445515603706.

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

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

Cooper, J. O., Heron, T. E., & Heward, W. L. (2007). *Applied behavior analysis* (2nd ed.). Upper Saddle River: Pearson Education, Inc.

Danziger, K. (1990). *Constructing the subject: Historical origins of psychological research.* Cambridge, UK: Cambridge University Press.

Davidson, P. O., & Costello, C. G. (Eds.). (1969). *N = 1: Experimental studies of single cases.* New York, NY: Van Nostrand Reinhold.

De, T. K., Michiels, B., Vlaeyen, J. W., & Onghena, P. (2017). Shiny SCDA. [Computer software]. Retrieved from https://ppw.kuleuven.be/mesrg/software-and-apps/shiny-scda.

de Jong, J., Vlaeyen, J., Onghena, P., Cuypers, C., den Hollander, M., & Ruijgrok, J. (2005). Reduction of pain-related fear in complex regional pain syndrome type I: The application of graded exposure in vivo. *Pain, 116*, 264–275. https://doi.org/10.1016/j.pain.2005.04.019.

Dugard, P. (2014). Randomization tests: A new gold standard? *Journal of Contextual Behavioral Science, 3*, 65–68. https://doi.org/10.1016/j.jcbs.2013.10.001.

Dugard, P., File, P., & Todman, J. (2011). *Single-case and small-N experimental designs.* New York, NY: Routledge Academic.

Dukes, W. F. (1965). N = 1. *Psychological Bulletin, 64*, 74–79. https://doi.org/10.1037/h0021964.

Edgington, E. S. (1967). Statistical inference from N = 1 experiments. *Journal of Psychology, 65*, 195–199. https://doi.org/10.1080/00223980.1967.10544864.

Edgington, E. S. (1975). Randomization tests for one-subject operant experiments. *Journal of Psychology, 90*, 57–68. https://doi.org/10.1080/00223980.1975.9923926.

Edgington, E. S. (1980a). Overcoming obstacles to single-subject experimentation. *Journal of Educational Statistics, 5*, 261–267. https://doi.org/10.3102/10769986005003261.

Edgington, E. S. (1980b). Validity of randomization tests for one-subject experiments. *Journal of Educational Statistics, 5*, 235–251. https://doi.org/10.2307/1164966.

Edgington, E. S. (1996). Randomized single-subject experimental designs. *Behaviour Research and Therapy, 34*, 567–574. https://doi.org/10.1016/0005-7967(96)00012-5.

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

Evans, J. J., Gast, D. L., Perdices, M., & Manolov, R. (2014). Single case experimental designs: Introduction to a special issue of Neuropsychological Rehabilitation. *Neuropsychological Rehabilitation, 24*, 305–314. https://doi.org/10.1080/09602011.2014.903198.

Fedorov, S. (2013). GetData graph digitizer. Retrieved from http://getdata-graph-digitizer.com/.

Ferron, J., & Foster-Johnson, L. (1998). Analyzing single-case data with visually guided randomization tests. *Behavior Research Methods, Instruments, & Computers, 30*, 698–706. https://doi.org/10.3758/BF03209489.

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

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

Ferron, J., & Onghena, P. (1996). The power of randomization tests for single-case phase designs. *The Journal of Experimental Education, 64*, 231–239. https://doi.org/10.1080/00220973.1996.9943805.

Ferron, J., & Ware, W. (1995). Analyzing single-case data: The power of randomization tests. *The Journal of Experimental Education, 63*, 167–178. https://doi.org/10.1080/00220973.1995.9943820.

Fisher, R. A. (1935). *The design of experiments.* Edinburgh, UK: Oliver & Boyd.

Fisher, A. J., Medaglia, J. D., & Jeronimus, B. F. (2018). A lack of group-to-individual generalizability is a threat to human subjects research. *PNAS: Proceedings of the National Academy of Sciences of the United States of America, 115*, 6106–6115. https://doi.org/10.1073/pnas.1711978115.

Guyatt, G. H., Haynes, R. B., Jaeschke, R. Z., Cook, D. J., Green, L., Naylor, C. D., et al. (2000). Users' guides to the medical literature: XXV. Evidence-based medicine: Principles for applying the users' guides to patient care. *Journal of the American Medical Association, 284*, 1290–1296. https://doi.org/10.1001/jama.284.10.1290.

Guyatt, G., Jaeschke, R., & McGinn, T. (2002). PART 2B1: Therapy and validity. N-of-1 randomized controlled trials. In G. Guyatt, D. Rennie, M. O. Meade, & D. J. Cook (Eds.). *Users' guides to the medical literature* (pp. 275–290). New York, NY: McGraw-Hill.

Hamaker, E. L. (2012). Why researchers should think "within-person": A paradigmatic rationale. In M. R. Mehl, & T. S. Conner (Eds.). *Handbook of research methods for studying daily life* (pp. 43–61). New York, NY: Guilford Press.

Hartmann, D. P., & Hall, R. V. (1976). The changing criterion design. *Journal of Applied Behavior Analysis, 9*, 527–532. https://doi.org/10.1901/jaba.1976.9-527.

Hersen, M., & Barlow, D. H. (1976). *Single case experimental designs: Strategies for studying behavior change.* New York, NY: Pergamon Press.

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

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

Heyvaert, M., & Onghena, P. (2014b). Randomization tests for single-case experiments: State of the art, state of the science, and state of the application. *Journal of Contextual Behavioral Science, 3*, 51–64. https://doi.org/10.1016/j.jcbs.2013.10.002.

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

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

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* (pp. 271–305). (3rd ed.). Boston, MA: Allyn & Bacon.

Howard, D., Best, W., & Nickels, L. (2015). Optimising the design of intervention studies: Critiques and ways forward. *Aphasiology, 29*, 526–562. https://doi.org/10.1080/02687038.2014.985884.

Howick, J., Chalmers, I., Glasziou, P., Greenhalgh, T., Heneghan, C., Liberati, A., et al. (2011). *The Oxford levels of evidence 2.* Oxford Centre for Evidence-Based Medicine. Retrieved from http://www.cebm.net/index.aspx?o=5653.

Kazdin, A. E. (1980). Obstacles in using randomization tests in single-case experimentation. *Journal of Educational Statistics, 5*, 253–260. https://doi.org/10.3102/10769986005003253.

Kazdin, A. E. (1982). *Single-case research designs: Methods for clinical and applied settings.* London, UK: Oxford University Press.

Kazdin, A. E. (1984). Statistical analyses for single-case experimental designs. In D. H. Barlow, & M. Hersen (Eds.). *Single case experimental designs: Strategies for studying behavior change* (pp. 258–324). (2nd ed.). New York, NY: Pergamon.

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

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole.

Klein, L. A., Houlihan, D., Vincent, J. L., & Panahon, C. J. (2017). Best practices in utilizing the changing criterion design. *Behavior Analyst in Practice, 10*, 52–61. https://doi.org/10.1007/s40617-014-0036-x.

Kratochwill, T. R. (Ed.). (1978). *Single subject research: Strategies for evaluating change.* New York, NY: Academic Press.

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

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

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

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

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

Ledford, J. R., & Gast, D. L. (Eds.). (2018). *Single case research methodology: Applications in special education and behavioral sciences*(3rd ed.). New York, NY: Routledge.

Lenz, A. S. (2015). Special issue editor's introduction: Using single-case research designs to demonstrate evidence for counseling practices. *Journal of Counseling and Development, 93*, 387–393. https://doi.org/10.1002/jcad.12036.

Levin, J. R., Evmenova, A. S., & Gafurov, B. S. (2014a). The single-case data-analysis ExPRT (excel package of randomization tests). In T. R. Kratochwill, & J. R. Levin (Eds.). *Single-case intervention research: Methodological and statistical advances* (pp. 185–219). Washington, DC: American Psychological Association.

Levin, J. R., Ferron, J. M., & Gafurov, B. S. (2014b). Improved randomization tests for a class of single-case intervention designs. *Journal of Modern Applied Statistical Methods, 13*, 2–52. https://doi.org/10.22237/jmasm/1414814460.

Levin, J. R., Ferron, J. M., & Gafurov, B. S. (2016). Comparison of randomization-test procedures for single-case multiple-baseline designs. *Developmental Neurorehabilitation,* 1–21. https://doi.org/10.1080/17518423.2016.1197708 July 1 [Epub ahead of print].

Levin, J. R., Ferron, J. M., & Gafurov, B. S. (2017). Additional comparisons of

randomization-test procedures for single-case multiple-baseline designs: Alternative effect types. *Journal of School Psychology, 63*, 13–34. https://doi.org/10.1016/j.jsp.2017.02.003.

Lillie, E. O., Patay, B., Diamant, J., Issell, B., Topol, E. J., & Schork, N. J. (2011). The n-of-1 clinical trial: The ultimate strategy for individualizing medicine? *Personalized Medicine, 8*, 161–173. https://doi.org/10.2217/pme.11.7.

Lundervold, D. A., & Belwood, M. F. (2000). The best kept secret in counseling: Single-case (N = 1) experimental designs. *Journal of Counseling and Development, 78*, 92–102. https://doi.org/10.1002/j.1556-6676.2000.tb02565.x.

Maggin, D. M., & Chafouleas, S. M. (2013). Introduction to the special series: Issues and advances of synthesizing single-case research. *Remedial and Special Education, 34*, 3–8. https://doi.org/10.1177/0741932512466269.

Maggin, D. M., Lane, K. L., & Pustejovsky, J. E. (2017). Introduction to the special issue on single-case systematic reviews and meta-analyses. *Remedial and Special Education, 38*, 323–330. https://doi.org/10.1177/0741932517717.

Manolov, R., Gast, D. L., Perdices, M., & Evans, J. J. (2014). Single-case experimental designs: Reflections on conduct and analysis. *Neuropsychological Rehabilitation, 24*, 634–660. https://doi.org/10.1080/09602011.2014.903199.

Manolov, R., & Moeyaert, M. (2017). Recommendations for choosing single-case data analytical techniques. *Behavior Therapy, 48*, 97–114. https://doi.org/10.1016/j.beth.2016.04.008.

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

McDougall, D. (2005). The range-bound changing criterion design. *Behavioral Interventions, 20*, 129–137. https://doi.org/10.1002/bin.189.

McDougall, D. (2013). Applying single-case design innovations to research in sport and exercise psychology. *Journal of Applied Sport Psychology, 25*, 33–45. https://doi.org/10.1080/10413200.2012.720640.

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

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement: Interdisciplinary Research and Perspectives, 2*, 201–218.

Morley, S. (2018). *Single case methods in clinical psychology: A practical guide.* London, UK: Routledge.

Nikles, J., & Mitchell, G. (Eds.). (2015). *The essential guide to N-of-1 trials in health.* Dordrecht, the Netherlands: Springer.

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

Onghena, P. (2018). Randomization and the randomization test: Two sides of the same coin. In V. Berger (Ed.). *Randomization, masking, and allocation concealment* (pp. 185–207). Boca Raton, FL: Chapman & Hall/CRC Press.

Onghena, P., & Edgington, E. S. (1994). Randomization tests for restricted alternating treatments designs. *Behaviour Research and Therapy, 32*, 783–786. https://doi.org/10.1016/0005-7967(94)90036-1.

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

Onghena, P., Maes, B., & Heyvaert, M. (in press). Mixed methods single case research: State of the art and future directions. Journal of Mixed Methods Research. http://doi.org/10.1177/1558689818789530.

Onghena, P., Michiels, B., Jamshidi, L., Moeyaert, M., & Van den Noortgate, W. (2018). One by one: Accumulating evidence by using meta-analytical procedures for single-case experiments. *Brain Impairment, 19*, 33–58. https://doi.org/10.1017/BrImp.2017.25.

Onghena, P., Vlaeyen, J., & de Jong, J. (2007). Randomized replicated single-case experiments: Treatment of pain-related fear by graded exposure in vivo. In S. Sawilowsky (Ed.). *Real data analysis* (pp. 387–396). Charlotte, NC: Information Age Publishing.

Parker, R. I., & Brossart, D. F. (2003). Evaluating single-case research data: A comparison of seven statistical methods. *Behavior Therapy, 34*, 189–211. https://doi.org/10.1016/S0005-7894(03)80013-8.

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

Parsonson, B. S., & Baer, D. M. (1978). The analysis and presentation of graphic data. In T. R. Kratochwill (Ed.). *Single-subject research: Strategies for evaluating change* (pp. 101–165). New York: Academic Press.

Pitman, E. J. G. (1937). Significance tests which may be applied to samples from any populations. *Supplement to the Journal of the Royal Statistical Society, 4*, 119–130. https://doi.org/10.2307/2984124.

Schork, N. J. (2015). Personalized medicine: Time for one-person trials. *Nature, 520*, 609–611. https://doi.org/10.1038/520609a.

Shadish, W. R. (2014). Analysis and meta-analysis of single-case designs: An introduction. *Journal of School Psychology, 52*, 109–122. https://doi.org/10.1016/j.jsp.2013.11.009.

Shadish, W. R., Kyse, E. N., & Rindskopf, D. M. (2013). Analyzing data from single-case designs using multilevel models: New applications and some agenda items for future research. *Psychological Methods, 18*, 385–405. https://doi.org/10.1037/a0032964.

Shamseer, L., Sampson, M., Bukutu, C., Schmid, C. H., Nikles, J., Tate, R., ... the CENT group (2015). CONSORT extension for reporting N-of-1 trials (CENT) 2015: Explanation and elaboration. *British Medical Journal, 350*, h1793. https://doi.org/10.1136/bmj/h1793.

Sidman, M. (1960). *Tactics of scientific research: Evaluating experimental data in psychology.* New York, NY: Basic Books.

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

Tate, R. L., Perdices, M., Rosenkoetter, U., McDonald, S., Togher, L., Shadish, W., ... Vohra, S. (2016b). The single-case reporting guideline in BEhavioural interventions (SCRIBE) 2016: Explanation and elaboration. *Archives of Scientific Psychology, 4*, 10–31. https://doi.org/10.1037/arc0000027.

Tate, R. L., Perdices, M., Rosenkoetter, U., Shadish, W., Vohra, S., Barlow, D. H., ... Wilson, B. (2016a). The Single-Case Reporting guideline in BEhavioural interventions (SCRIBE) 2016 statement. *Archives of Scientific Psychology, 4*, 1–9. https://doi.org/10.1037/arc0000026.

ter Kuile, M., Bulté, I., Weijenborg, P., Beekman, A., Melles, R., & Onghena, P. (2009). Therapist-aided exposure for women with lifelong vaginismus: A replicated single-case design. *Journal of Consulting and Clinical Psychology, 77*, 149–159. https://doi.org/10.1037/a0014273.

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. https://doi.org/10.1080/09602011.2013.824383.

The University of Sydney (2018). The SCRIBE 2016 statement. Retrieved from http://sydney.edu.au/medicine/research/scribe/statement.php.

Valsiner, J. (1986). Where is the individual subject in scientific psychology? In J. Valsiner (Ed.). *The individual subject and scientific psychology* (pp. 1–16). New York, NY: Plenum.

Van den Noortgate, W., & Onghena, P. (2003a). Combining single-case experimental data using hierarchical linear models. *School Psychology Quarterly, 18*, 325–346. https://doi.org/10.1521/scpq.18.3.325.22577.

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

Vlaeyen, J., de Jong, J., Geilen, M., Heuts, P., & van Breukelen, G. (2001). Graded exposure in vivo in the treatment of pain-related fear: A replicated single-case experimental design in four patients with chronic low back pain. *Behaviour Research and Therapy, 39*, 151–166. https://doi.org/10.1016/S0005-7967(99)00174-6.

Vohra, S. (2016). N-of-1 trials to enhance patient outcomes: Identifying effective therapies and reducing harms, one patient at a time. *Journal of Clinical Epidemiology, 76*, 6–8. https://doi.org/10.1016/j.jclinepi.2016.03.028.

Vohra, S., Shamseer, L., Sampson, M., Bukutu, C., Schmid, C. H., Tate, R., ... the CENT group (2015). CONSORT extension for reporting N-of-1 trials (CENT) 2015 Statement. *British Medical Journal, 350*, h1738. https://doi.org/10.1136/bmj/h1738.

Welch, B. L. (1937). On the *z*-test in randomized blocks and Latin squares. *Biometrika, 29*, 21–52. https://doi.org/10.2307/2332405.

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