Please cite this article as:

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

Randomization tests for single-case experiments: State of the art, state of the science, and state of the application

Abstract: A single-case experimental design is a research design that can be used to evaluate the effect of an intervention for a single entity. There are two important schedules to include randomization into the design of single-case experiments: phase designs and alternation designs. We present these two schedules and provide a detailed example for each schedule. For both examples, we illustrate the use of a free software package that assists researchers in designing and analyzing single-case experiments using randomization tests. Furthermore, we discuss several additions (simultaneous and sequential replication designs; meta-analysis of single-case experimental studies) and alternatives (statistical and visual analysis methods).
Keywords: single-subject experimental design; randomization test; phase design; alternation design; statistical software

1. State of the art

1.1 Single-case experiments

A single-case experiment (SCE) is an experiment that can be used to evaluate the effect of an independent variable for a single entity, for example a single patient, a single therapist-patient dyad, or a single family. An experimental approach is used: the independent variable is manipulated by the experimenter. The dependent variable is measured repeatedly for this entity under different levels of the independent variable. For example, in research on interventions for reducing challenging behavior among persons with autism, a simple SCE involves one person with autism showing certain challenging behavior (e.g., aggressive episodes) that is treated by an intervention (e.g., a behavioral intervention). The number of aggressive episodes of that person is repeatedly measured before (= baseline phase), during (= intervention phase), and after (= withdrawal phase) the behavioral intervention, during a certain period of time (e.g., eight weeks). Based on the difference between the number of aggressive episodes per day (i.e., the dependent variable) in the baseline phase versus in the intervention phase (i.e., the independent variable), the effect of the behavioral intervention on the challenging behavior can be determined for this person.

SCEs have a long history in the behavioral sciences, with pioneers like Ebbinghaus, Fechner, Stratton, and Wundt in the 19th century, and Pavlov, Sidman, and Skinner in the 20th century (Barlow, Nock, & Hersen, 2009; Blampied, 1999; Kratochwill & Mace, 1984).

We refer the reader interested in the historical and philosophical foundations of SCEs to the chapter of Ittenbach and Lawhead (1996). SCE research is often applied in several subdisciplines of the behavioral sciences, such as clinical psychology, counseling psychology, neuropsychology, school psychology, psychopharmacology, social work, education, and special education (e.g., Barlow et al., 2009; Dugard, File, & Todman, 2012; Heyvaert, Maes, Van den Noortgate, Kuppens, & Onghena, 2012; Horner et al., 2005; Kratochwill & Levin, 1992, 2010; Maggin, O'Keeffe, & Johnson, 2011; Perdices & Tate, 2009; Rapoff & Stark, 2008; Zhan & Ottenbacher, 2001).

There are several reasons to explain the growing interest in and popularity of SCE research in the behavioral sciences. We list three basic reasons and five pragmatic reasons. A first basic reason is that the focus of SCE research on the individual case parallels the care for the individual patient in applied clinical settings (cf. Hayes, 1981). SCEs render results that are easily understood by clinicians who work at the level of individual patients (Rapoff & Stark, 2008). A second basic reason is that sometimes one SCE is sufficient to refute a hypothesis, or to confirm the presence of a phenomenon (Edelson, 1985; Onghena, 2005). A third basic reason concerns the growing importance of evidence-based practice, accountability, and of evaluating interventions at the level of the individual participant (Barlow, Hayes, & Nelson, 1985; Horner et al., 2005; Zhan & Ottenbacher, 2001).

A first pragmatic reason is that SCE research is one of the only eligible design options if rare or unique conditions are involved (e.g., a patient with a rare psychological disorder). A second pragmatic reason is that it is sometimes sufficient to use SCE research (or a few replications) when the between-case variability is very small, for instance for very homogeneous groups. In that case, it can be far more interesting to study the within-case variability than the between-case variability. A third pragmatic reason for the growing interest in SCE research in the behavioral sciences is its feasibility and flexibility (Hacker, 1980; McReynolds & Thompson, 1986). A fourth pragmatic reason for the growing impact of SCE research is the present-day availability of several methods for the design and analysis of SCE research (cf. Barlow et al., 2009; Dugard et al., 2012; Edgington & Onghena, 2007, pp. 225-259; Franklin, Allison, & Gorman, 1996; Kazdin, 2011; Parker, Vannest, & Davis, 2011), as well as software for the design and analysis of SCE research (e.g., Bulté & Onghena, 2008, 2009, 2012; Koehler & Levin, 2000; Van den Noortgate & Onghena, 2003a, 2007). A fifth pragmatic reason for the popularity of SCE research is its small-scale design: a small-scale design is less harmful and less costly than a large-scale design.

1.2 Randomization

In group-comparison studies, randomization concerns the random assignment of participants to control and treatment groups. In SCEs, randomization concerns the random assignment of measurement times to baseline and treatment conditions. Applying randomization increases the methodological quality of a study, whether it is a group-comparison study or an SCE. For group-comparison studies, the 'randomization' element of the randomized controlled trial (RCT) design led to its status of the gold standard for evaluating the efficacy of treatments. Accordingly, reporting tools for RCTs and tools for evaluating the methodological quality of RCTs include items on randomization (e.g., Altman et al., 2001; Maher, Sherrington, Herbert, Moseley, & Elkins, 2003). Concerning SCEs as well, the advantages of including randomization in the design of the experiment are described in several books and articles (e.g., Barlow et al., 2009; Dugard et al., 2012; Edgington & Onghena, 2007; Kazdin, 2011; Kratochwill & Levin, 1992, 2010). Accordingly, we notice that some recently developed SCE reporting tools and tools for evaluating the methodological quality of SCEs include items on randomization too (e.g., Romeiser-Logan, Hickman, Harris, & Heriza, 2008; Task Force on Evidence-Based Interventions in School Psychology, 2003).

We argue that randomization is as important for SCEs as it is for group-comparison studies. A researcher conducting a *nonrandomized* SCE has to be very careful when attributing changes in the outcomes to changes in the treatment conditions, because it is possible that the observed response trend of the participant might have been there *without* any treatment manipulation. However, in a *randomized* SCE the random assignment of measurement times to baseline and treatment conditions provides control for sources of bias (Edgington, 1987, 1996). For SCEs the randomization process can render control over - both known and unknown - confounding variables that are time-related, such as maturation effects (Onghena, 2005). As such, randomization can increase the methodological quality of an SCE by strengthening the internal validity of SCEs (Edgington & Onghena, 2007; Kratochwill & Levin, 2010; Onghena & Edgington, 2005). However, there are limits on the control established, depending on the SCE design and the applied randomization procedure. When measurement times are randomly assigned to baseline and treatment conditions (cf. **2.3**), intervention assignment is unrelated to time. When only the intervention start points are randomly determined (cf. **2.1**), there is still some relationship between treatment assignment and time.

1.3 Randomization tests

Randomization tests (RTs) are statistical significance tests based on the random assignment of experimental units to treatments (Edgington & Onghena, 2007). They are used to test hypotheses about treatment effects. Using RTs increases the methodological quality of an SCE by improving the statistical conclusion validity of SCEs (Edgington & Onghena, 2007; Kratochwill & Levin, 2010; Onghena & Edgington, 2005). Let us look in more detail at the steps involved for the RT (cf. Edgington & Onghena, 2007; Ferron & Ware, 1995). As a prerequisite, the randomization method requires that a researcher can design his experimental study so that it involves random assignment. A priori, all possible random assignments are recorded. Randomly one of these assignments is chosen: our actual experiment will follow this assignment. Then, the researcher chooses an appropriate test statistic, runs the experiment, collects the data, and calculates the test statistic based on the collected data. We then look at all possible random assignments that were recorded at the beginning of our study: each of these assignments involved a different way of dividing the data. We calculate the chosen test statistic for each of these assignments. Based on this, we can determine the statistical significance of our test statistic: we look where our obtained test statistic falls within the distribution of all possible test statistic values. The *p* value of the RT is calculated as the proportion of possible test statistic values that is as extreme, or even more extreme, than the value of the test statistic based on the collected data.

2. Randomization tests for single-case experiments

There are two important schedules to include randomization into the design of SCEs: phase designs and alternation designs. In the former, the moment of phase change is randomly determined. In the latter, the treatment alternation is randomly determined. We now consecutively present these two schedules (**2.1**; **2.3**), and describe a detailed example for each schedule (**2.2**; **2.4**).

2.1 Randomization tests for single-case phase designs

When using phase designs, all measurement times are divided into phases and several consecutive measurements are taken in each phase (Edgington, 1975, 1980; Onghena, 1992). The AB design is the most basic phase design: the A phase is the baseline or control phase, and the B phase is the intervention phase. There exist several extensions and variations of the AB phase design: ABA designs (i.e., reversal and withdrawal designs), BAB designs, ABAB designs, ABABAB designs, and so on (Barlow et al., 2009, pp. 135-161; Kazdin, 2011, pp.

121-143). In **Figure 1**, three examples of single-case phase designs are presented. It is also possible that more than one treatment is evaluated by means of a single-case design. With the B phase representing the consecutive measurements taken under the first treatment, and the C phase representing the consecutive measurements taken under the second treatment, several other extensions and variations are possible, such as ABACA designs, ABCB designs, and so on (Barlow et al., 2009, pp. 162-175).

Insert Figure 1 about here

In all these phase designs, the sequence of the phases is fixed before the start of the actual experiment. The incorporation of randomization in phase designs concerns the moment of phase change. For instance, in an AB design the randomization concerns the moment when the intervention (B phase) starts. In an ABAB design the randomization concerns the moment when the first intervention phase, the second baseline phase, and the second intervention phase start. When designing a randomized phase SCE, three features should be decided a priori. First, the total number of available measurement times should be decided. For example, if we want to measure the number of aggressive episodes per day for a period of eight weeks, there are 56 measurement times. Second, the number of phases should be decided a priori. For example, if we want to use an ABAB design to study the effect of the behavioral intervention on the challenging behavior, there are four consecutive phases involved. Third, the minimum lengths of the phases can be decided a priori. In a *completely randomized* phase design, we do *not* decide a priori the minimum lengths of the phases. However, in that case it is possible that very few measurement times are assigned to a phase, for instance that only one measurement occasion is assigned to the first intervention phase in our ABAB design. In these completely randomized phase designs, it is even possible that zero measurement times are assigned to a phase. Because researchers usually do not want that there are too few, or even zero, measurement times for one phase, they often apply *restricted randomized* phase designs. In these designs, the minimum number of measurement times for each phase is decided a priori. So, the total number of available measurement times (our first feature; e.g., 56 measurement times) is divided into phases (our second feature; e.g., ABAB design), with a fixed minimum number of measurement times for each phase (e.g., minimum 5 measurement times for each phase).

The larger the set of potential randomizations generated by the SCE design, the smaller the smallest possible *p* value will be for the phase design. For example, a randomized

AB design with eight measurement times and with at least two measurement times in each phase, implies the following five design possibilities: AABBBBBB, AAABBBBB, AAAABBBB, AAAAABBB, and AAAAAABB. However, a randomized ABAB design with 56 measurement times and with at least 5 measurement times in each phase implies 9139 design possibilities (Onghena, 1992). In the former example the smallest possible *p* value is .2 (1/5), while the smallest possible *p* value is about .0001 (1/9139) in the latter example.

2.2 Randomization tests for single-case phase designs: An example

We present two examples for single-case RTs: one for phase designs (**2.2**) and one for alternation designs (**2.4**). For each example, we provide a step-by-step guide. Bulté and Onghena (2008, 2009, 2012) developed a software package in R that helps researchers to design and analyze SCEs: the SCDA package. This package provides a Graphical User Interface for the Single-Case Visual Analysis (SCVA), Single-Case Randomization Test (SCRT) and Single-Case Meta-Analysis (SCMA) packages. The steps and the R code for designing and analyzing single-case phase and alternation designs are described in more detail in the article of Bulté and Onghena (2008).

As an illustration for phase designs, we conduct an RT for the data set from Lorimer, Simpson, Myles, and Ganz (2002) on the use of social stories as a preventative behavioral intervention in a home setting with a 5-year-old boy with autism (Gregg). The data are displayed in **Figure 2**.

Insert Figure 2 about here

The first step concerns the choice of the design. Lorimer et al. (2002) used an ABAB phase design.

Second, the null hypothesis and the alternative hypothesis must be formulated, and an appropriate test statistic should be selected. The *null hypothesis* says that there is no treatment effect of the social story intervention: the responses of the participant are independent of the condition (‘baseline’ versus ‘social stories’) under which they are observed. The *alternative hypothesis* can be formulated in a directional, one-tailed manner or in a nondirectional, two-tailed manner. For this example, the alternative hypothesis can be formulated in a one-tailed manner because the direction of the effect can be specified: Lorimer et al. (2002) expected *positive* effects of the social story intervention on the frequency of the interrupting verbalizations of the participant. They expected a *smaller* frequency of disruptive vocalizing

in the intervention phases, compared to the baseline phases. When possible, it is better to use the one-tailed option, because the RT will then be more powerful.

Next, an appropriate *test statistic* should be selected: the test statistic has to be chosen in accordance to the kind of effects the researcher expects or predicts. One of the advantages of RTs is that the choice of the test statistic is not confined to statistics for which the sampling distribution has been derived or tabulated (Onghena & Edgington, 2005). As such, the researcher is offered maximal flexibility in defining the test statistic. A popular choice for the test statistic of the RT is the difference between the (sums of) means for the baseline and treatment phases. Other statistics that can be used to test for other kinds of predicted effects are for instance differences between the medians, the ratio of the variances, and differences between the slopes (Onghena & Edgington, 2005). Furthermore, it is also possible to use an effect size index as a test statistic for the RT (cf. Heyvaert & Onghena, 2013). As such, not only the (non)randomness of an intervention effect, but also the direction and magnitude of this effect can be determined. Different effect size indices might be appropriate for different SCE designs and data patterns (see e.g., Gage & Lewis, 2013; Manolov et al., 2011, for a discussion). One group of SCE effect size measures that are closely related to visual analysis are the 'nonoverlap statistics', such as PND (percent of nonoverlapping data), PZD (percentage of zero data points), PEM (percent of data points exceeding the median), IRD (improvement rate difference), and PAND (percent of all nonoverlapping data) (see e.g., Parker et al., 2011, for a review of several nonoverlap statistics). A second group of SCE effect size measures is based on the standardized mean differences (SMD) effect sizes used for group-comparison designs (e.g., Cohen's *d*, Glass's Δ, Hedges' *g*). A third group are regression-based effect size measures: regression techniques are used to estimate the effect size for an SCE by taking a trend into account. In **3.3** we discuss the advantages and disadvantages for these three groups of SCE effect size indices. Based on the expectations and predictions of Lorimer et al. (2002), we apply the difference between the sums of means for the baseline and treatment phases as test statistic for this example (cf. Onghena, 1992): $T = (\overline{A_1} + \overline{A_2}) - (\overline{B_1} + \overline{B_2})$.

Third, the level of significance, α, and the number of measurement times should be determined. For our example the significance level is .05 and the total number of measurement times is 24.

Fourth, the randomization schedule and the selection of the assignment should be determined. For the ABAB design in our example, this means that we should specify the minimum number of measurement times for each phase. Suppose that this minimum was set

to three. To determine how many possible data arrangements there are for this specific design, we can use the R code developed by Bulté and Onghena (2008)[1]. In the menu SCDA, we successively click on 'SCRT', 'DESIGN YOUR EXPERIMENT', and 'NUMBER OF POSSIBLE ASSIGNMENTS'. We select the design type 'ABAB PHASE DESIGN', set 'NUMBER OF OBSERVATIONS' to 24, and set the limitation (i.e., the minimum number of observations per phase) to 3. The result of this computation is 455, which means that for an ABAB design with 24 measurement times and a minimum of 3 measurement times per phase, the total number of possible data arrangements is 455. The data set from Lorimer et al. (2002) represents 1 of these 455 possible data arrangements, with the following sequence: A1 A1 A1 A1 A1 A1 A1 B1 B1 B1 B1 B1 B1 B1 A2 A2 A2 B2 B2 B2 B2 B2 B2 B2.

Fifth, the data should be collected and the observed test statistic is computed. The data from Lorimer et al. (2002) are shown in **Table 1**. In order to use the R functions described further in this article, we should create a text (.txt) file containing the data. The data file should consist of two columns, that are separated by a tab. The first column should contain the subsequent condition labels (independent variable): A1 A1 A1 A1 A1 A1 A1 B1 B1 B1 B1 B1 B1 B1 A2 A2 A2 B2 B2 B2 B2 B2 B2 B2. The second column should contain the obtained scores for each measurement time (dependent variable): 4 6 3 4 5 4 7 4 2 2 2 1 2 1 3 7 7 3 2 2 1 0 1 0. As such, each row represents one measurement occasion. The rows and columns must not be labeled. We save our text file as 'ABAB.txt'.

Insert Table 1 about here

This text file can be used to generate a graphical presentation of the collected data, as in **Figure 2**. We can do this in the menu SCDA. We successively click on 'SCVA' and 'GRAPHICAL DISPLAY'. We select 'USE DATA FROM TEXT FILE' and subsequently select our text file 'ABAB.txt'. We select the design type 'ABAB PHASE DESIGN'. This results in **Figure 2**.

In order to calculate the observed test statistic from the obtained raw data, we again use the menu SCDA. We successively click on 'SCRT', 'ANALYZE YOUR DATA', and 'OBSERVED TEST STATISTIC'. We select 'USE DATA FROM TEXT FILE' and

[1] R can be downloaded freely from the CRAN Web site (http://cran.r-project.org/). We used the R version 2.15.1. After opening R, we successively click on 'PACKAGES', 'LOAD PACKAGE', and 'RCMDRPLUGIN.SCDA'. Afterwards, the R COMMANDER appears containing the additional menu SCDA.

subsequently select our text file ‘ABAB.txt’. We select the design type ‘ABAB PHASE DESIGN’ and the statistic ‘AA-BB’. Our observed test statistic is 7.095238.

Sixth, we can construct the randomization distribution (see **1.3**). RTs take all possible combinations - given the applied randomization procedure - of the collected data into consideration (i.e., 455 possible data assignments for our example). To test the null hypothesis that there is no treatment effect of the social story intervention, the test statistic has to be calculated for each of these 455 possible data arrangements. In order to do that, all 24 scores are kept fixed (i.e., 4 6 3 4 5 4 7 4 2 2 2 1 2 1 3 7 7 3 2 2 1 0 1 0), but the phase transitions (i.e., from A1 to B1, from B1 to A2, and from A2 to B2) are randomly determined taking into account the minimum of 3 measurement times for each phase. Afterwards, the 455 test statistics are sorted in ascending order, which forms the randomization distribution under the null hypothesis. In order to construct the randomization distribution, we can use the menu SCDA. We successively click on ‘SCRT’, ‘ANALYZE YOUR DATA’, and ‘RANDOMIZATION DISTRIBUTION’. We select ‘USE DATA FROM TEXT FILE’ and subsequently select our text file ‘ABAB.txt’. We select the design type ‘ABAB PHASE DESIGN’, the statistic ‘AA-BB’, and the limitation 3. We select ‘SYSTEMATIC RANDOMIZATION DISTRIBUTION’, since the number of possible data assignments for our example is not that large, and so it is feasible for R to compute the test statistics for all possible data divisions. Afterwards, the 455 test statistics appear in the output window.

Seventh, we can assess the statistical significance of the outcome (*p* value). We do this by positioning the observed test statistic in the randomization distribution: the proportion of test statistics in the randomization distribution that exceeds or equals our observed test statistic (i.e., 7.095238) corresponds with the RT’s *p* value. We reject the null hypothesis, and accept the alternative hypothesis, when the *p* value is smaller than or equal to the level of significance α (.05 for our example). In order to determine the *p* value, we again use the menu SCDA. We successively click on ‘SCRT’, ‘ANALYZE YOUR DATA’, and ‘P VALUE’. We select ‘USE DATA FROM TEXT FILE’ and select our text file ‘ABAB.txt’. We select the design type ‘ABAB PHASE DESIGN’, the statistic ‘AA-BB’, and the limitation 3. We again select ‘SYSTEMATIC RANDOMIZATION DISTRIBUTION’. This gives a *p* value of 0.01318681, which is smaller than the predetermined significance level of .05. Accordingly, the null hypothesis that the responses of the participant are independent of the condition (‘baseline’ versus ‘social stories intervention’) under which they are observed, can be rejected. The interrupting verbalizations of the participant are significantly less frequent in the social story intervention phases than in the baseline phases.

For this example, we illustrated how to use the R Commander for analyzing the SCE data of Lorimer et al. (2002). However, one could also use the R code in **Appendix 1**: the same results are obtained, without using the R Commander.

2.3 Randomization tests for single-case alternation designs

Researchers can design an alternation SCE if quick and repeated alternation of treatments is possible for one participant (Barlow & Hayes, 1979). Any level of the independent variable could be present at each measurement occasion in alternation designs (Edgington, 1967; Onghena & Edgington, 1994). Alternation designs can be used to compare the effects of an intervention condition to the effects of a control, placebo, or baseline condition. However, alternation designs can also be used to compare the effects of one intervention to the effects of a second intervention. In **Figure 3** an example of a single-case alternation design is presented.

Insert Figure 3 about here

For phase designs, we differentiated between *completely randomized* designs and *restricted randomized* designs. For alternation designs we make a distinction between *completely randomized* designs, *randomized block* designs, and *alternating treatments* designs (Barlow et al., 2009, pp. 243-270; Onghena, 2005). When the data are classified according to more than one independent variable, factorial designs can be used for analyzing the SCE data (cf. Edgington & Onghena, 2007, pp. 89-108).

For alternation designs, the RT is based on the random ordering of the levels of the independent variable (Onghena & Edgington, 2005). In *completely randomized* alternation designs, the treatment sequence is randomly determined by taking into account two features. The first is the number of levels of the independent variable: for instance, two levels for an alternation design studying the effects of one treatment condition versus a control condition. The second is the number of measurement times for each level. If you design an alternation SCE with 2 levels (A and B), with 3 measurement times for each level, then there are 20 possible sequences: AAABBB, AABABB, AABBAB, AABBBA, ABAABB, ABABAB, ABABBA, ABBAAB, ABBABA, ABBBAA, BAAABB, BAABAB, BAABBA, BABAAB, BABABA, BABBAA, BBAAAB, BBAABA, BBABAA, and BBBAAA. Complete randomization implies a random selection of 1 sequence among these 20 possible sequences.

However, sometimes certain assignments of a completely randomized alternation design could be undesirable. It is plausible that we are not interested in sequences with three successive identical conditions, such as AAABBB and ABBBAA. In that case, other alternation designs can be applied. We present two alternatives. First, if we would pair each AB sequence from the previous example, and then randomly determine the order of the two members for each pair, we would design a *randomized block* alternation SCE. For our example, this results in eight possible sequences: ABABAB, ABABBA, ABBAAB, ABBABA, BABABA, BABAAB, BAABBA, and BAABAB. So, the six sequences with three successive identical treatments are no longer generated as possible sequences (i.e., AAABBB, AABBBA, ABBBAA, BAAABB, BBAAAB, and BBBAAA).

Nevertheless, this randomized block alternation design is rather restrictive: six other sequences as well are no longer generated as possible sequences: AABABB, AABBAB, ABAABB, BABBAA, BBAABA, and BBABAA. Hence, we present a second alternative to the completely randomized alternation design: a randomized version of the *alternating treatments* design along with an algorithm to enumerate and randomly sample the set of all acceptable sequences (Onghena & Edgington, 1994). This second alternative offers us all 14 sequences we are interested in, and only excludes the 6 sequences with 3 successive identical treatments.

When discussing phase designs (**2.1**), we pointed at considering the smallest possible *p* value. For alternation designs too, the larger the set of potential randomizations generated by the SCE design, the smaller the smallest possible *p* value will be. For our example, 14 potential sequences were generated by the alternating treatments design, while only 8 potential sequences were generated by the randomized block design. Using the same number of measurement times, randomized alternating treatments designs may guarantee smaller possible *p* values than randomized block designs.

2.4 Randomization tests for single-case alternation designs: An example

As an illustration for alternation designs, we conduct an RT for the data set from Van Laarhoven, Kraus, Karpman, Nizzi, and Valentino (2010, p. 202) on the comparison of the effectiveness of video prompting and picture prompting when used as antecedents for teaching daily living skills to a 13-year-old boy with autism, Marvin. The data for Marvin are displayed in **Figure 4**. The seven steps are identical to the seven steps described for the ABAB phase design example (**2.2**). In what follows, we will illustrate how to use the R Commander for analyzing the SCE data of Van Laarhoven et al. (2010). However, one could

also use the R code in **Appendix 2**: the same results are obtained, without using the R Commander.

Insert Figure 4 about here

First, we choose an appropriate design. Van Laarhoven et al. (2010) used an alternating treatments design.

Second, the null hypothesis and the alternative hypothesis should be formulated, and an appropriate test statistic should be selected. There is an important difference between the data from Lorimer et al. (2002) and Van Laarhoven et al. (2010). Lorimer et al. (2002) measured the frequency of interrupting vocalizations (i.e., undesirable behavior as dependent variable). Accordingly, we expected a *smaller* frequency of the dependent variable in the intervention condition ('social stories'), compared to the baseline condition. Van Laarhoven et al. (2010) measured the percentage of independent correct responses in task analyses (i.e., desirable behavior as dependent variable). For the study of Van Laarhoven et al. (2010) the *null hypothesis* says that there is no difference between the effects of video prompting and picture prompting on Marvin's percentage of independent correct responses. We formulate the *alternative hypothesis* in a directional, one-tailed manner: we expect a *higher* percentage of independent correct responses in the video prompting condition, compared to the picture prompting condition. Based on the expectations and predictions of Van Laarhoven et al. (2010), we use the difference between the means for the two conditions: $T = \bar{B} - \bar{A}$, with B = 'video prompting' and A = 'picture prompting', as test statistic for this example (cf. **2.2** for a discussion of alternative test statistics).

Third, the number of measurement times and the level of significance should be determined. For our example the number of instructional sessions equals 10, and the significance level is .05.

Fourth, the randomization schedule and the selection of the assignment should be determined. For the alternating treatments design in our example, this means that we should specify the maximum number of consecutive administrations of the same condition (i.e., A or B). Suppose that this maximum was set to two for each condition. We call this the limitation. In the menu SCDA, we successively click on 'SCRT', 'DESIGN YOUR EXPERIMENT', and 'NUMBER OF POSSIBLE ASSIGNMENTS'. We select the design type 'ALTERNATING TREATMENTS DESIGN', set 'NUMBER OF OBSERVATIONS' to 10, and set the limitation to 2. The result of this computation is 84, which means that for an

alternating treatments design with 2 levels, with 10 measurement times and a maximum of 2 consecutive administrations of the same condition, the total number of possible data arrangements is 84. The data set from Van Laarhoven et al. (2010) represents 1 of these 84 possible data arrangements, with the following sequence: BAABBABABA.

Fifth, the data should be collected and the observed test statistic should be computed. The raw data collected by Van Laarhoven et al. (2010) for Marvin are shown in **Table 2**. In order to use the R functions described further in this article, we should create a text (.txt) file containing the data. The data file should consist of two columns, that are separated by a tab. The first column should contain the subsequent condition labels (independent variable): BAABBABABA. The second column should contain the obtained scores for each measurement time (dependent variable): 80 55 55 90 90 85 90 85 100 100. As such, each row represents one measurement occasion. The rows and columns must not be labeled. We save our text file as 'Alternating.txt'.

Insert Table 2 about here

This text file can be used to generate a graphical presentation of the collected data, as in **Figure 4**. We can do this in the menu SCDA. We successively click on 'SCVA' and 'GRAPHICAL DISPLAY'. We select 'USE DATA FROM TEXT FILE' and subsequently select our text file 'Alternating.txt'. We select the design type 'ALTERNATING TREATMENTS DESIGN'. This results in **Figure 4**.

In order to calculate the observed test statistic from the obtained raw data, we again use the menu SCDA. We successively click on 'SCRT', 'ANALYZE YOUR DATA', and 'OBSERVED TEST STATISTIC'. We select 'USE DATA FROM TEXT FILE' and subsequently select our text file 'Alternating.txt'. We select the design type 'ALTERNATING TREATMENTS DESIGN' and the statistic 'B-A'. Our observed test statistic is 14.

Sixth, we can construct the randomization distribution (see **1.3**). To test the null hypothesis that there is no difference between the effects of video prompting and picture prompting on Marvin's percentage of independent correct responses, the test statistic has to be calculated for each of the 84 possible data arrangements. In order to do that, all 10 scores are kept fixed (i.e., 80 55 55 90 90 85 90 85 100 100), but the conditions (i.e., A and B) assigned to these 10 measurement times are randomly shuffled according to the predefined possible orderings (i.e., an alternating treatments design with a maximum of 2 consecutive administrations of the same condition). Afterwards, the 84 test statistics are sorted in

ascending order, which forms the randomization distribution under the null hypothesis. **Figure 5** shows an histogram of the randomization distribution for the data set from Van Laarhoven et al. (2010).

In order to construct the randomization distribution, we can use the menu SCDA. We successively click on 'SCRT', 'ANALYZE YOUR DATA', and 'RANDOMIZATION DISTRIBUTION'. We select 'USE DATA FROM TEXT FILE' and subsequently select our text file 'Alternating.txt'. We select the design type 'ALTERNATING TREATMENTS DESIGN', the statistic 'B-A', and the limitation 2. We select 'SYSTEMATIC RANDOMIZATION DISTRIBUTION'. Afterwards, the 84 test statistics appear in the output window.

Insert Figure 5 about here

Seventh, we can assess the statistical significance of the outcome. We reject the null hypothesis, and accept the alternative hypothesis, when the *p* value is smaller than or equal to the level of significance α (.05 for our example). In order to determine the *p* value, we again use the menu SCDA. We successively click on 'SCRT', 'ANALYZE YOUR DATA', and 'P VALUE'. We select 'USE DATA FROM TEXT FILE' and select our text file 'Alternating.txt'. We select the design type 'ALTERNATING TREATMENTS DESIGN', the statistic 'B-A', and the limitation 2. We again select 'SYSTEMATIC RANDOMIZATION DISTRIBUTION'. This gives a *p* value of 0.04761905. This value is smaller than the predetermined significance level of .05. Accordingly, the null hypothesis that there is no difference between the effects of video prompting and picture prompting on Marvin's percentage of independent correct responses, can be rejected. The percentage of independent correct responses in the video prompting condition is significantly higher than in the picture prompting condition.

3. Additions and alternatives

Some important aims of contextual behavioral science (CBS) are the understanding of behavior and the promotion of human growth and development. SCE designs can be used by CBS researchers for the measurement of processes of change and behavioral outcomes (cf. Hayes, Barnes-Holmes, & Wilson, 2012; Vilardaga, Hayes, Levin, & Muto, 2009). As such, CBS researchers can use SCE designs for testing CBS theories and for studying the complexity of human behavior at the level of the individual.

In **3.1** and **3.2** we discuss how the SCE designs presented in **2.1** and **2.3** can be replicated and combined in a meta-analysis. Until a few decades ago, the lack of generally accepted rules for drawing generalization inferences from SCE data was considered one of the major drawbacks of SCEs (Kennedy, 1979). This contrasted with the well-established statistical methods developed to estimate the generalizability of the findings for group-comparison studies. For this reason, group-comparison studies were preferred over SCEs by many behavioral researchers. However, during the last 30 years several ways to demonstrate and test the external validity of SCE findings have been studied. Accordingly, nowadays different methods to include replication in the SCE design are applied in the behavioral sciences. Based on the temporal dimension, we differentiate between *simultaneous replication designs* and *sequential replication designs* (cf. Onghena & Edgington, 2005). CBS researchers are encouraged to replicate their analyses applied at the level of the individual in order to test the external validity of their findings (cf. Hayes et al., 2012; Vilardaga et al., 2009). *Meta-analysis techniques* can be used to systematically compare and synthesize the results of multiple SCE studies. Advantages of meta-analyses are that reasons for heterogeneity can be identified and new hypotheses can be generated about particular subgroups; that new perspectives and frameworks can be generated that transcend the retrieved SCE studies; that large amounts of information can be assimilated quickly by practitioners, researchers, and policymakers through consulting these meta-analyses; and that meta-analyses may reduce the delay between research discoveries and implementation of effective strategies in practice (Greenhalgh, 1997).

Afterwards, we discuss alternatives to RTs for the analysis of SCE data. We present some statistical (**3.3**) and visual (**3.4**) analysis methods, and discuss their strengths and weaknesses. Based on **3.3** and **3.4**, we advocate in **3.5** the combined use of visual analysis methods, RTs, effect size measures, and time series analysis. In **3.6** we emphasize the difference between clinical/practical and statistical significance.

3.1 Simultaneous replication designs

In simultaneous replication designs, the replications are implemented concurrently. In **Figure 6** an example of a simultaneous replication design is presented: a multiple baseline design (MBD; Barlow et al., 2009, pp. 201-241; Kazdin, 2011, pp. 144-166). The MBD is often applied by behavioral researchers. In an MBD across participants, several participants are included, and an AB phase design is carried out for each separate participant with different starting points for the intervention phase. The MBD involves simultaneous data

collection across multiple participants. In the example presented in **Figure 6**, four participants are involved. For each of them an AB phase design is implemented and there are different moments of phase change for each participant. The phase change for these 4 participants occurs respectively at measurement time 5, 7, 9 and 11.

Insert Figure 6 about here

A major advantage of simultaneous replication designs is that the effect of historical confounding variables can be effectively assessed and ruled out because the data are collected concurrently. In an MBD corresponding changes across the data series of all included participants indicate a potential historical confounding variable, and, hence, historical confounding variables can be ruled out when there are no changes that correspond across the data series (Christ, 2007). If a phase change is implemented for one participant (e.g., start of the intervention phase at measurement time 5 for participant 1) and produces a change in the level of the dependent variable for that participant, while little or no change is observed for the other included participants (e.g., at measurement time 5 participants 2, 3, and 4 are still in the baseline phase), then it is less likely that other external events are responsible for the change in the dependent variable, and more likely that the independent variable is responsible for the change in the dependent variable.

How can we introduce randomization for simultaneous replication designs? We can do this by separately implementing the randomization schedules in the several phase or alternation designs for all included participants (Onghena, 1992). Accordingly, the RT is based on the randomization in each of the phase or alternation designs separately and uses a multivariate statistic to test the null hypothesis that there is no effect for any of the patients. A multivariate test statistic that can be used is the mean difference between the phase averages, $T = \sum_{i=1}^{k} \overline{A}_i - \overline{B}_i$ with k equal to the number of participants (Onghena & Edgington, 2005). The steps and R code for designing and analyzing MBDs are described in detail in the article of Bulté and Onghena (2009). They are parallel to the examples described in **2.2** and **2.4**.

The randomization strategy for simultaneous replication designs presented in this section is described by Marascuilo and Busk (1988) and Onghena (1992). Alternative randomization schedules for MBDs are described by Wampold and Worsham (1986) and Koehler and Levin (1998). In order to avoid that exactly the same randomization schedule is selected for the included participants, the latter authors proposed to impose a restriction that ensures temporal staggering: 'regulated randomization'. This option is also included in free

software packages for designing and analyzing SCEs. For instance, in the SCDA package different possible starting points for the intervention phase can be inserted for each participant when designing randomized MBDs, in order to avoid overlap between the possible starting points of the different participants.

3.2 Meta-analysis of SCE studies and sequential replication designs

Within the evidence-based practice movement, researchers and practitioners increasingly rely on meta-analyses to render guidelines for best practice (Beretvas & Chung, 2008; Shadish & Rindskopf, 2007). Some important benefits of meta-analytic research over single primary studies are: a higher statistical power to detect effects, more accurate effect size estimations, the ability to make more convincing generalizations to a larger population, the ability to identify sources of heterogeneity and to test moderators to explain detected between-study variation, and the ability to detect biases (e.g., publication bias). Meta-analytic methods for group-comparison studies described in general handbooks cannot simply be applied to SCE data, because effects for group-comparison studies are generally based on the comparison between two independent groups with scores within each condition coming from different participants, while effects for SCEs are based on scores from the same participant over time (Beretvas & Chung, 2008; Busk & Serlin, 1992; Van den Noortgate & Onghena, 2008). Accordingly, specific procedures have been developed for the meta-analysis of SCE studies. We make a distinction between two groups of procedures: procedures for combining *p* values and procedures for combining effect sizes.

Methods for *combining p values* of SCEs (cf. step 7 in **2.2** and **2.4**) are for instance the Stouffer method (see Rosenthal, 1978), the additive method (Edgington, 1972a; Onghena & Edgington, 2005), the multiplicative method (Edgington, 1972a; Pesarin, 2001; Rosenthal, 1978), the normal curve method for situations with four or more *p* values (Edgington, 1972b), and the iterative procedure for combining *p* values (Pesarin, 2001; Pesarin & Salmaso, 2010). With the SCDA package it is possible to combine *p* values of SCEs: in the menu SCDA, you successively click on ‘SCMA’ and ‘COMBINE P VALUES’ (Bulté, Van Den Noortgate, & Onghena, 2010, May). The SCDA package so far includes the additive and multiplicative combining approach.

The second group of procedures focuses on combining *effect size indices*. In **2.2** we mentioned three groups of effect size indices: nonoverlap statistics (e.g., PND, PZD, PEM, IRD, PAND), SMD effect sizes, and regression-based effect size measures. The potential strengths and weaknesses for these three groups of SCE effect size indices will be discussed

in **3.3**. With the SCDA package it is possible to calculate some of these indices: in the menu SCDA, you successively click on 'SCMA' and 'CALCULATE EFFECT SIZES' (Bulté et al., 2010, May). The SCDA package so far includes the following three effect sizes: SMD, PND, and PEM. The simplest method for combining effect sizes of SCEs is taking the (weighted) average of the effect sizes of all included SCEs in a meta-analysis. More advanced methods for combining effect sizes of SCEs are for instance Busk and Serlin's (1992) three approaches (the choice for an approach depends on the assumptions made on the variability and distribution of the data), and the hierarchical linear model or multilevel model proposed by Van den Noortgate and Onghena (2003a, 2003b, 2008).

In **3.1** we discussed how to design and analyze simultaneous replication designs using RTs. For sequential replication designs the replications over the included participants are carried out consecutively (cf. **Figure 7** for an example). Both simultaneous and sequential replication designs offer protection against historical confounding variables (cf. internal validity), and both have different participants working on different randomized schedules. The main differences concern the set-up of the replication study and the analysis of the collected data. Regarding the set-up, researchers involved in simultaneous replication designs have to manipulate the experimental variables and monitor the behavior of all included participants in the same period of time. Although the intervention is applied sequentially across the participants, the dependent variables have to be measured for all participants in the same period of time, which can increase the workload of the researchers. In sequential replication designs the included sub-studies are carried out consecutively, i.e. without demanding the participants to be monitored at the same time, thus reducing the workload of the researchers at one moment in time. Concerning the analysis of the collected data, we discussed in **3.1** how the RT for simultaneous replication designs uses a multivariate statistic to test the null hypothesis that there is no effect for any of the patients. Since the replicated SCEs in sequential replication designs can be considered as separate studies, they can be analyzed by the abovementioned meta-analytic procedures (i.e., procedures for combining *p* values and procedures for combining effect sizes).

Insert Figure 7 about here

3.3 Statistical analysis of single-case data

The SCE literature differentiates between two groups of methods for the analysis and interpretation of SCE data: statistical analysis (**3.3**) and visual analysis (**3.4**). A problem with

SCE data is that they often contain serial dependencies. This means that scores at one measurement point in SCE data series are predictive of scores at another measurement point in these data series. Serial dependency can be measured by calculating an autocorrelation coefficient. Hence when analyzing SCE data it is advised to use an analysis method that addresses autocorrelation.

Statistical methods traditionally used for analyzing group-comparison studies have been applied to analyze SCE data in the past (e.g., Gentile, Roden, & Klein, 1972). However, traditional parametric tests such as the *t*-test and *F*-test are often not appropriate to analyze SCE data because (a) these data often violate assumptions of normality, (b) these tests assume that observations in each phase are independent while SCE data are often autocorrelated, and (c) these tests are based on means and variances alone and are insensitive to trends that occur within a phase (Houle, 2009).

Accordingly, several nonparametric tests have been proposed as alternatives, such as the Kruskal-Wallis test (1952), the Mann-Whitney *U* test (1947), and RTs. The present article described how to conduct RTs for SCE data and discussed their benefits. Like the other nonparametric tests, RTs are distribution-free tests that do not demand that the SCE data are serially independent. This makes them suitable options for the analysis of SCE data. An advantage of RTs to the Wilcoxon-Mann-Whitney test and the Kruskal-Wallis test, is that the *p* value can be derived without degrading the observed scores to ranks (Onghena & Edgington, 2005). However, there are some potential drawbacks of RTs. We discussed that measurement times should be randomly assigned to baseline and treatment conditions (‘randomized designs’) in order to use RTs. Because this randomization has to be done before the data are collected, RTs seem incompatible with response-guided experimentation. Nonetheless, it is possible to combine randomization and response-guided experimentation procedures in SCEs. A researcher can decide to make the manipulation of the conditions only partially dependent on the data: by using such a ‘restricted random assignment’ valid significance determination by the RT procedure is still possible (Edgington, 1980). For example, a researcher conducting an AB phase design can a priori decide not to introduce the experimental treatment until after the baseline data show stability (cf. response-guided experimentation), but he can supplement this procedure with the random selection of the moment when the intervention phase starts, *after* baseline stability has been attained, thereby allowing for the valid use of an RT (Edgington, 1975). Ferron, Foster-Johnson, and Kromrey (2003) and Levin, Ferron, and Kratochwill (2011) compared simulation results for the application of RT procedures in non-randomized designs to the application in randomized

designs: for the former designs the test often failed to have Type I error rates that matched the nominal level, while for the latter designs Type I error rates were well controlled. When applied to a non-randomized design the test may be better referred to as a 'restricted permutation test'.

A second possible limitation of RTs concerns statistical power. The statistical power varies depending on the design, on the test statistic, on the level of autocorrelation, on the magnitude of the effect size, on the number of observations, and on the number of possible assignments or permutations (e.g., Dugard et al., 2012; Ferron & Onghena, 1996; Ferron & Sentovich, 2002; Ferron & Ware, 1995; Levin, Ferron, & Kratochwill, 2012; Levin, Lall, & Kratochwill, 2011; Manolov, Solanas, Bulté, & Onghena, 2010). For instance, although the power of individual AB designs usually is very low, the combination of three or more AB designs in an MBD raises the power considerably.

In the two examples discussed in this article, the number of observations is rather limited: there are 24 observations in the first example (cf. **2.2**) and 10 in the second example (cf. **2.4**). Although we - for didactical reasons - work with small examples for this article, we want to stress that having a considerable number of observations under each condition is critical in developing a sound SCE design. However, the number of observations and the number of possible assignments do certainly not cover the whole story, e.g., an AB design with 25 potential start points does not have more power than a four participant MBD with a randomly assigned intervention order resulting in 24 possible assignments. More potential randomizations lead to smaller potential *p* values (cf. supra), but not necessarily to more power.

As an alternative to RTs, time series analysis is considered a valuable statistical method for the analysis and interpretation of SCE data because of its ability to handle serial dependency (Box & Jenkins, 1970; Glass, Willson, & Gottman, 1975; Gottman & Glass, 1978; Hartmann et al., 1980; Jones, Vaught, & Reid, 1975; Jones, Vaught, & Weinrott, 1977). With time-series analysis one can determine whether an experimental manipulation or a clinical intervention has produced a reliable change in temporally ordered scores (Hartmann et al., 1980). Unfortunately, because time series analysis requires a large number of measurement times per phase, it is limited in its practical use. Another major disadvantage is that the success of the time series analysis approach depends on the fitting of a suitable model to describe the nature of the autocorrelation in the SCE data, while several problems are reported with the process of identifying a suitable model (Houle, 2009). There is no single way of doing time series analysis that can be recommended as widely applicable across most

designs and data sets: there are multiple options that yield different results (Kazdin, 2011). As such, time series analysis is a technically complex approach that might be difficult to grasp for clinicians and applied researchers.

Additionally, effect sizes can be used to quantify the amount of behavior change between baseline/control and intervention phases in SCEs. In **2.2** we briefly introduced three groups of SCE effect sizes: nonoverlap statistics, SMD effect sizes, and regression-based effect size measures. General advantages of the first group of effect sizes, the nonoverlap statistics, are their accordance with visual analysis and their ease to calculate and interpret. However, for now each available nonoverlap statistic is equipped to adequately address only a relatively narrow spectrum of SCE designs and different disadvantages are noted for different nonoverlap statistics, such as deficient performance in the presence of data outliers in the baseline phase, insensitivity to data trends and variability in the data, and insensitivity to differences in the magnitude of effect (Maggin et al., 2011; Smith, 2012; Wolery, Busick, Reichow, & Barton, 2010). For the second group of effect size measures, it is important to note that for computing SMDs for group-comparison designs the *variation between groups* is used, while the *within-case variation* is used for computing SMDs for SCE designs. Due to this computational difference, the obtained SCE effect sizes are not interpretively equivalent to SMDs for group-comparison designs. An exception is the recently developed SMD effect size for SCEs of Hedges, Pustejovsky, and Shadish (2012): it is directly comparable with Cohen's *d*. Advantages of this second group of SCE effect size measures are their ease of use and their familiarity to applied researchers; other drawbacks are that SMDs were not developed to contend with autocorrelated data and that SMDs are insensitive to trends (Maggin et al., 2011; Smith, 2012). For the third group, regression techniques are used to estimate the effect size by taking a trend into account. Examples are the piecewise regression approach of Center, Skiba, and Casey (1985-1986), the approach of White, Rusch, Kazdin, and Hartmann (1989), the approach of Allison and Gorman (1993), and hierarchical linear models (e.g., Van den Noortgate & Onghena, 2003a, 2003b). Advantages of these regression-based effect sizes are their ability to account for linear or nonlinear trends in the data as well as for dependent error structures within the SCE data (Maggin et al., 2011; Van den Noortgate & Onghena, 2003b). A possible limitation of regression-based effect sizes is that most applied researchers are not familiar with their calculation and interpretation, and that the required statistical techniques may be far more technically challenging than the procedures required to calculate and interpret the first two groups of SCE effect sizes.

3.4 Visual analysis of single-case data

The second group of methods used to analyze and interpret SCE data are visual analysis methods. These methods aim at reaching a judgment about the reliability or consistency of intervention effects by visually examining graphed data (Kazdin, 1982, p. 232). In contrast with RTs, the visual analysis method is not aimed at directly testing the null hypothesis. The method visually examines the graphical display of the SCE data, with the scores on the dependent variable on the vertical *y*-axis and the measurement times on the horizontal *x*-axis. Visual analysis is implemented by judging the extent of changes in level, variability, and trend, and the latency of change evident across phases, and whether the changes are consistent with the requirements of the particular design (Kazdin, 2011). When the changes in level and/or variability are in the desired direction and when they are immediate, readily discernible, and maintained over time, it is concluded that the changes in behavior across phases result from the implemented treatment and are indicative of improvement (Busse, Kratochwill, & Elliott, 1995). Demonstration of a functional relationship between the independent and dependent variable is compromised when there is a long latency between manipulation of the independent variable and change in the dependent variable, when level changes across conditions are small and/or similar to changes within conditions, and when trends do not conform to those predicted following introduction or manipulation of the independent variable (Horner et al., 2005).

The SCDA package includes a basic plotting function to display data from SCEs as well as tools to facilitate the use of three major interpretative principles of visual analysis: central location and level, variability, and trend across baseline and intervention conditions (Bulté & Onghena, 2012). The major strengths of visual analysis are that it is quick, easy, and inexpensive to use in applied clinical settings, and that it is widely accepted and understood. Accordingly, visual analysis is still the most popular method for the analysis and interpretation of SCE data in the behavioral sciences (Busk & Marascuilo, 1992; Kazdin, 2011; Kratochwill & Brody, 1978).

Although visual analysis is very important when it comes to presenting the results in a paper, the method brings along some serious drawbacks when used as a decision procedure, especially in borderline cases. First, research has shown that the average interrater agreement, and thus the reliability or consistency, of the visual analysis of SCE data is low (e.g., DeProspero & Cohen, 1979; Harbst, Ottenbacher, & Harris, 1991; Matyas & Greenwood, 1990a, 1990b; Ottenbacher, 1990; Park, Marascuilo, & Gaylord-Ross, 1990). Second, the accuracy of visual analysis is questioned because of the inflated type I error rates, i.e.

concluding a treatment produced an effect when in fact the results could be due to chance (e.g., Borckardt, Murphy, Nash, & Shaw, 2004; Ferron & Jones, 2006; Matyas & Greenwood, 1990a). Additionally, Kazdin (1982, 2011) describes several situations in which visual analysis is inadequate: (1) in SCEs without a stable baseline (i.e., trend in the data), (2) in SCEs with increased intrasubject variability, (3) when a new treatment is studied that does not result in immediate and obvious treatment effects, or when treatment effects are weak or ambiguous, (4) when small but important and reliable changes in the performance of the individual subjects occur, and (5) when it is necessary to statistically control for extraneous factors when the SCE is implemented in the natural environment.

Summarizing, we advise all SCE researchers publishing in the domain of CBS to use visual analysis to present their results in a paper, but since the method brings along some serious drawbacks when used alone as a decision tool, we advise to complement it with statistical approaches in the analysis and interpretation of SCE data. Visual analysis of SCEs is particularly prone to error when the treatment effects are small and unstable (Busse et al., 1995). In these instances, additionally conducting statistical analyses may be especially useful (Kazdin, 2011).

3.5 Combined use of visual analysis methods, randomization tests, effect size measures, and time series analysis

We argue that the combined use of visual analysis methods, RTs, effect size measures, and time series analysis methods is necessary for a comprehensive and reliable analysis and interpretation of SCE data. It usually makes sense, and often is a necessary step, to use RTs in order to rule out the null hypothesis that there is no differential effect of the conditions (Onghena & Edgington, 2005). Addressing the drawbacks of visual analysis of SCE data (cf. **3.4**), statistical significance tests produce consistent results that are independent of the performer of the analysis, they can be applied with unstable baseline data and with increased intrasubject variability, and they have the ability to detect small but systematic treatment effects that might be ignored by visual analysis (Kazdin, 2011; Nourbakhshs & Ottenbacher, 1994).

However, to facilitate the interpretation of the results of an RT, and to emphasize the difference between the nonrandomness of an effect and the magnitude of an effect, it is recommended to supplement the results of the RT with an effect size measure (Onghena & Edgington, 2005). Nowadays several leading organizations stress the importance of reporting effect sizes for primary outcomes (e.g., American Psychological Association, 2010, p. 34;

Kratochwill et al., 2010; Task Force on Evidence-Based Interventions in School Psychology, 2003). Since effect size measures can be used as test statistics in RTs, the two can be combined when analyzing SCEs: the effect size for the observed SCE data can be measured and this value can be located in a distribution of values that were equally likely if the null hypothesis is true (cf. Heyvaert & Onghena, 2013).

Nor statistical significance tests nor effect size measures provide the breadth and integrated nature of holistic visual analysis methods that simultaneously consider mean or median level shifts, trends and trend differences, abrupt or gradual behavior changes at intervention onset, data variability within and across phases, differences in trend line intercepts at intervention points, curvilinear progress, lag or delay in response to intervention, and so on (Parker & Hagan-Burke, 2007; Parker, Vannest, & Brown, 2009). Additionally, the methods and results of statistical significance tests and effect size measures can be difficult to grasp and interpret for clinicians and applied researchers, while the visual inspection of SCE data is more intuitively appealing and easier to understand.

Since serial dependency may have an influence on the results and conclusions of visual analyses and RTs, an SCE report should also address the issue of autocorrelation by means of, for instance, time series analysis of the SCE data (cf. **3.3**). Besides being useful in quantifying autocorrelation, statistical models like time series analyses or multilevel models may complement the RT and visual analysis in quantifying the size of the treatment effect, the amount the treatment effect varies over time within a participant, and the amount that the effect varies across participants (e.g., Van den Noortgate & Onghena, 2007).

Like we mentioned in **3.3** and **3.4**, each method has his drawbacks. Applied together the methods can control and compensate for each other's drawbacks, and provide a more reliable analysis and interpretation of the SCE data.

3.6 Clinical and statistical significance

Finally, we find it important to emphasize the difference between clinical/practical and statistical significance. Statistical significance tests like RTs (cf. **3.3**) only concern statistical significance and should not be used as indicators of clinical significance: a statistically significant difference in outcome measures between the baseline and treatment phases does not necessarily mean that the treatment has had a clinically significant impact, particularly if the effect size is small (Onghena & Edgington, 2005; Perdices & Tate, 2009). Although visual analysis methods (cf. **3.4**) and effect sizes (cf. **3.3**) can be useful when assessing the clinical significance of behavior change, the construct of clinical significance is broader than that

(Allison & Gorman, 1993). Clinical significance refers to the improvement in the dependent variable as well as to the practical importance of the effect of an intervention: whether it makes a 'real' difference to the patient and/or to others with whom the patient interacts in everyday life (Kazdin, 1999; Perdices & Tate, 2009). Apart from the reliability of behavior change (cf. statistical significance tests) and the magnitude of experimental effects (cf. effect size measures), the importance of the change and the impact on patient functioning adds critical dimensions (Kazdin, 1999). Measures that can be used to assess clinical significance are for instance: falling within a normative range, departure from dysfunctional behavior, no longer meeting diagnostic criteria, quality of life, and risk-benefit contours (cf. Kazdin, 2011; Schulz et al., 2002; Shakespeare, Gebski, Veness, & Simes, 2001). Accordingly, in clinical single-case practice it should be recommended to use a two-fold criterion for determining improvement in a patient, based on both statistical reliability and clinical significance (Jacobson, Follette, & Revenstorf, 1984).

4. Conclusion

SCE designs can be used by contextual behavioral science researchers for testing theories and for studying the complexity of human behavior at the level of the individual. Including randomization in SCE studies can considerably increase the methodological quality of the studies. In this article, we discussed two important schedules to include randomization into the design of SCEs and provided an empirical example for each schedule. Recently, free software has been developed to assist contextual behavioral science researchers in designing and analyzing randomized SCEs, using randomization tests. We illustrated the use of a free software package in R for our two examples and provided a step-by-step guide for the design and analysis of the two SCEs. Furthermore, we discussed important additions to the basic SCE design: simultaneous and sequential replication designs and meta-analysis of SCE studies can help contextual behavioral science researchers to draw generalization inferences from SCE data. Next, we discussed alternatives to randomization tests for the analysis of SCE data. We described several statistical and visual analysis methods, and discussed their strengths and weaknesses. Based on this discussion, our advice for contextual behavioral science researchers is to combine visual analysis methods, effect size measures, and randomization tests in the analysis of SCE data: applied together the methods can control for each other's drawbacks, and provide a reliable analysis of the SCE data.

References

Allison, D. B., & Gorman, B. S. (1993). Calculating effect sizes for meta-analysis: The case of the single case. *Behaviour Research and Therapy, 31,* 621-631. doi:10.1016/0005-7967(93)90115-B

Altman, D. G., Schulz, K. F., Moher, D., Egger, M., Davidoff, F., Elbourne, D., et al. (2001). The revised CONSORT statement for reporting randomized trials: Explanation and elaboration. *Annals of Internal Medicine, 134,* 663-694.

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

Barlow, D. H. & Hayes, S. C. (1979). Alternating treatments design: One strategy for comparing the effects of two treatments in a single subject. *Journal of Applied Behavior Analysis, 12,* 199-210. doi:10.1901/jaba.1979.12-199

Barlow, D. H., Hayes, S. C., & Nelson, R. O. (1985). *The scientist practitioner: Research and accountability in clinical and educational settings*. Oxford, UK: Pergamon.

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

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. doi:10.1080/17489530802446302

Blampied, N. M. (1999). A legacy neglected: Restating the case for single-case research in cognitive-behaviour therapy. *Behaviour Change, 16,* 89-104. doi:10.1375/bech.16.2.89

Borckardt, J. J., Murphy, M. D., Nash, M. R., & Shaw, D. (2004). An empirical examination of visual analysis procedures for clinical practice evaluation. *Journal of Social Service Research, 30,* 55-73. doi:10.1300/J079v30n03_04

Box, G. E. P., & Jenkins, G. M. (1970). *Time series analysis: Forecasting and control*. San Francisco, CA: Holden Day.

Bulté, I., & Onghena, P. (2008). An R package for single-case randomization tests. *Behavior Research Methods, 40,* 467-478. doi: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. doi: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. doi:10.1027/1614-2241/a000042

Bulté, I., Van Den Noortgate, W., & Onghena, P. (2010, May). *An R package for the nonparametric meta-analysis of small-n educational studies*. Poster presented at the Annual Meeting of the American Educational Research Association (AERA), Denver, Colorado (USA).

Busk, P. L., & Marascuilo, L. A. (1992). Statistical analysis in single-case research: Issues, procedures, and recommendations, with applications to multiple behaviors. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 159-185). Hillsdale, NJ: Lawrence 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, NJ: Lawrence Erlbaum.

Busse, R. T., Kratochwill, T. R., & Elliott, S. N. (1995). Meta-analysis for single-case consultation outcomes: Applications to research and practice. *Journal of School Psychology, 33*, 269-285. doi:10.1016/0022-4405(95)00014-D

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. doi:10.1177/002246698501900404

Christ, T. J. (2007). Experimental control and threats to internal validity of concurrent and nonconcurrent multiple baseline designs. *Psychology in the Schools, 44*, 451-459. doi:10.1002/pits.20237

DeProspero, A., & Cohen, S. (1979). Inconsistent visual analysis of intrasubject data. *Journal of Applied Behavior Analysis, 12*, 285-296. doi:10.1901/jaba.1979.12-573.

Dugard, P., File, P., & Todman, J. (2012). *Single-case and small-n experimental designs: A practical guide to randomization tests* (2nd ed.). New York, NY: Routledge Academic.

Edelson, M. (1985). *Hypothesis and evidence in psychoanalysis*. Chicago, IL: University of Chicago Press.

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

Edgington, E. S. (1972a). An additive method for combining probability values from independent experiments. *Journal of Psychology, 80*, 351-363. doi:10.1080/00223980.1972.9924813

Edgington, E. S. (1972b). A normal curve method for combining probability values from independent experiments. *Journal of Psychology, 82*, 85-89. doi:10.1080/00223980.1972.9916971

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

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

Edgington, E. S. (1987). Randomized single-subject experiments and statistical tests. *Journal of Counseling Psychology, 34,* 437-442. doi:10.1037/0022-0167.34.4.437

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

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

Ferron, J., Foster-Johnson, L., & Kromrey, J. (2003). The functioning of single-case randomization tests with and without random assignment. *Journal of Experimental Education, 71,* 267-288. doi:10.1080/00220970309602066

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

Ferron, J., & Onghena, P. (1996). The power of randomization tests for single-case phase designs. *Journal of Experimental Education, 64,* 231-239.

Ferron, J., & Sentovich, C. (2002). Statistical power of randomization tests used with multiple-baseline designs. *Journal of Experimental Education, 70,* 165-178. doi:10.1080/00220970209599504

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

Franklin, R. D., Allison, D. B., & Gorman, B. S. (Eds.) (1996). *Design and analysis of single-case research*. Mahwah, NJ: Lawrence Erlbaum.

Gage, N. A., & Lewis, T. J. (2013). Analysis of effect for single-case design research. *Journal of Applied Sport Psychology, 25,* 46-60. doi:10.1080/10413200.2012.660673

Gentile, J. R., Roden, A. H., & Klein, R. D. (1972). An analysis-of-variance model for the intrasubject replication design. *Journal of Applied Behavior Analysis, 5,* 193-198. doi:10.1901/jaba.1972.5-193

Glass, G. V., Willson, V. L., & Gottman, J. M. (1975). *Design and analysis of time-series experiments*. Boulder, CO: Colorado Associated University Press.

Gottman, J. M., & Glass, G. V. (1978). Analysis of interrupted time series analysis. In T. R. Kratochwill (Ed.), *Single subject research*. New York, NY: Academic Press.

Greenhalgh, T. (1997). Papers that summarise other papers (systematic reviews and meta-analyses). *British Medical Journal, 315,* 672–675. doi:10.1136/bmj.315.7109.672

Hacker, B (1980). Single subject research strategies in occupational therapy, Part 1. *American Journal of Occupational Therapy, 34,* 103-108.

Harbst, K. B., Ottenbacher, K. J., & Harris, S. R. (1991). Interrater reliability of therapists' judgments of graphed data. *Physical Therapy, 71,* 107-115.

Hartmann, D. P., Gottman, J. H., Jones, R. R., Gardner, W., Kazdin, A. E., & Vaught, R. S. (1980). Interrupted time-series analysis and its application to behavioral data. *Journal of Applied Behavior Analysis, 13,* 543-559. doi:10.1901/jaba.1980.13-543

Hayes, S. C. (1981). Single case experimental design and empirical clinical practice. *Journal of Consulting and Clinical Psychology, 49,* 193-211. doi:10.1037/0022-006X.49.2.193

Hayes, S. C., Barnes-Holmes, D., & Wilson, K. G. (2012). Contextual Behavioral Science: Creating a science more adequate to the challenge of the human condition. *Journal of Contextual Behavioral Science, 1,* 1-16.

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. doi:10.1002/jrsm.1052

Heyvaert, M., Maes, B., Van den Noortgate, W., Kuppens, S., & Onghena, P. (2012). A multilevel meta-analysis of single-case and small-n research on interventions for reducing challenging behavior in persons with intellectual disabilities. *Research in Developmental Disabilities, 33,* 766-780. doi:10.1016/j.ridd.2011.10.010

Heyvaert, M., & Onghena, P. (2013). Analysis of single-case data: Randomization tests for measures of effect size. *Neuropsychological Rehabilitation*. doi:10.1080/09602011.2013.818564

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.

Houle, T. T. (2009). Statistical analyses for single-case experimental designs. In D. H. Barlow, M. K. Nock, & M. Hersen (Eds.), *Single case experimental designs: Strategies for studying behavior change* (3rd ed.) (pp. 271-305). Boston, MA: Allyn & Bacon.

Ittenbach, R. F., & Lawhead, W. F. (1996). Historical and philosophical foundations of single-case research. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 13-39). Mahwah, NJ: Lawrence Erlbaum.

Jacobson, N. S., Follette, W. C., & Revenstorf, D. (1984). Psychotherapy outcome research: Methods for reporting variability and evaluating clinical significance. *Behaviour Therapy, 15,* 336-352. doi:10.1016/S0005-7894(84)80002-7

Jones, R. R., Vaught, R. S., & Reid, J. B. (1975). Time series analysis as a substitute for single subject analysis of variance designs. In G. R. Patterson, I. M. Mark, J. D. Matarazzo, R. A. Myuers, G. E. Schwartz, & H. H. Strapp (Eds.), *Behavior change*. Chicago, IL: Aldine.

Jones, R. R., Vaught, R. S., & Weinrott, M. (1977). Time series analysis in operant research. *Journal of Applied Behavior Analysis, 10,* 151-166. doi:10.1901/jaba.1977.10-151

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

Kazdin, A. E. (1999). The meanings and measurement of clinical significance. *Journal of Consulting and Clinical Psychology, 67,* 332-339. doi:10.1037/0022-006X.67.3.332

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

Kennedy, M. M. (1979). Generalizing from single case studies. *Evaluation Review, 3,* 661-678. doi:10.1177/0193841X7900300409

Koehler, J., & Levin, J. (1998). Regulated randomization: A potentially sharper analytical tool for the multiple-baseline design. *Psychological Methods, 2,* 206-217. doi:10.1037/1082-989X.3.2.206

Koehler, J., & Levin, J. (2000). RegRand: Statistical software for the multiple-baseline design. *Behavior Research Methods, Instruments, & Computers, 32,* 367-371. doi:10.3758/BF03207807

Kratochwill, T. R., & Brody, G. H. (1978). Single subject designs: A perspective on the controversy over employing statistical inference and implications for research and training in behavior modification. *Behavior Modification, 2,* 291-307. doi:10.1177/014544557823001

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 http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf.

Kratochwill, T. R., & Levin, J. R. (Eds.). (1992). *Single-case research design and analysis: New directions for psychology and education*. Hillsdale, NJ: Lawrence Erlbaum.

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

Kratochwill, T. R., & Mace, F. C. (1984). Time-series designs in psychotherapy research. In M. Hersen, L. Michelson, & A. Bellack (Eds.), *Issues in psychotherapy research* (pp. 171-226). New York, NY: Plenum Press.

Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association, 47,* 583-621. doi:10.1080/01621459.1952.10483441

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. doi:10.1016/j.jsp.2012.05.001

Levin, J. R., Lall, V. F., & Kratochwill, T. R. (2011). Extensions of a versatile randomization test for assessing single-case intervention effects. *Journal of School Psychology, 49,* 55-79. doi:10.1016/j.jsp.2010.09.002

Lorimer, P. A., Simpson, R. L., Myles, B. S., & Ganz, J. B. (2002). The use of social stories as a preventative behavioral intervention in a home setting with a child with autism. *Journal of Positive Behavior Interventions, 4,* 53-60. doi:10.1177/109830070200400109

Maggin, D. M., O'Keeffe, B. V., & Johnson, A. H. (2011). A quantitative synthesis of methodology in the meta-analysis of single-subject research for students with disabilities: 1985-2009. *Exceptionality, 19,* 109-135. doi:10.1080/09362835.2011.565725

Maggin, D. M., Swaminathan, H., Rogers, H. J., O'Keeffe, B. V., Sugai, G., & Horner, R. H. (2011). A generalized least squares regression approach for computing effect sizes in single-case research: Application examples. *Journal of School Psychology, 49,* 301-321. doi:10.1016/j.jsp.2011.03.004

Maher, C. G., Sherrington, C., Herbert, R. D., Moseley, A. M., & Elkins, M. (2003). Reliability of the PEDro scale for rating quality of randomized controlled trials. *Physical Therapy, 83,* 713-721.

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. doi:10.1214/aoms/1177730491

Manolov, R., Solanas, A., Bulté, I., & Onghena, P. (2010). Data-division-specific robustness and power of randomization tests for ABAB designs. *Journal of Experimental Education, 78,* 191-214. doi:10.1080/00220970903292827

Manolov, R., Solanas, A., Sierra, V., & Evans, J. J. (2011). Choosing among techniques for quantifying single-case intervention effectiveness. *Behavior Therapy, 42,* 533-545.

Marascuilo, L., & Busk, P. (1988). Combining statistics for multiple-baseline AB and replicated ABAB designs across subjects. *Behavioral Assessment, 10,* 1-28.

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

Matyas, T. A., & Greenwood, K. M. (1990b). The effect of serial dependence on visual judgment of single-case charts: An addendum. *American Journal of Occupational Therapy, 10,* 308-320.

McReynolds, L. V., & Thompson, C. K. (1986). Flexibility of single-subject experimental designs. Part I: Review of the basics of single-subject designs. *Journal of Speech and Hearing Disorders, 51,* 194-203.

Nourbakhshs, M. R., & Ottenbacher, K. J. (1994). The statistical analysis of single-subject data: A comparative examination. *Physical Therapy, 74,* 768-776.

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

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

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

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

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

Park, H., Marascuilo, L., & Gaylord-Ross, R. (1990). Visual inspection and statistical analysis of single-case designs. *Journal of Experimental Education, 58,* 311-320.

Parker, R. I., & Hagan-Burke, S. (2007). Useful effect size interpretations for single case research. *Behavior Therapy, 38,* 95-105. doi:10.1016/j.beth.2006.05.002

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. (2011). Effect size in single-case research: A review of nine nonoverlap techniques. *Behavior Modification, 35,* 303-322. doi:10.1177/0145445511399147

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

Pesarin, F. (2001). *Multivariate permutation tests*. New York, NY: Wiley.

Pesarin, F., & Salmaso, L. (2010). *Permutation tests for complex data: Theory, applications and software*. Chichester, UK: Wiley.

Rapoff, M., & Stark, L. (2008). Editorial: Journal of Pediatric Psychology - Statement of purpose: Section on single-subject studies. *Journal of Pediatric Psychology, 33,* 16-21. doi:10.1093/jpepsy/jsm101

Romeiser-Logan, R. L., Hickman, R. R., Harris, S. R., & Heriza, C. B. (2008). Single-subject research design: Recommendations for levels of evidence and quality rating. *Developmental Medicine & Child Neurology, 50,* 99-103. doi:10.1111/j.1469-8749.2007.02005.x

Rosenthal, R. (1978). Combining results of independent studies. *Psychological Bulletin, 85,* 185-193. doi:10.1037//0033-2909.85.1.185

Schulz, R., O'Brien, A., Czaja, S., Ory, M., Norris, R., Martire, L. M., et al. (2002). Dementia caregiver intervention research: In search of clinical significance. *Gerontologist, 42*, 589-602.

Shadish, W. R., & Rindskopf, D. M. (2007). Methods for evidence-based practice: Quantitative synthesis of single-subject designs. *New Directions for Evaluation, 113,* 95-109. doi:10.1002/ev.217

Shakespeare, T. P., Gebski, V. J., Veness, M. J., & Simes, J. (2001). Improving interpretation of clinical studies by use of confidence levels, clinical significance curves, and risk-benefit contours. *Lancet, 357,* 1349-1353. doi:10.1016/S0140-6736(00)04522-0

Smith, J. D. (2012). Single-case experimental designs: A systematic review of published research and current standards. *Psychological Methods, 17,* 510-550. doi:10.1037/a0029312µ

Task Force on Evidence-Based Interventions in School Psychology (2003). *Procedural and coding manual for review of evidence-based interventions*. Retrieved from http://www.indiana.edu/~ebi/documents/_workingfiles/EBImanual1.pdf

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

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

Van den Noortgate, W., & Onghena, P. (2007). The aggregation of single-case results using hierarchical linear models. *The Behavior Analyst Today, 8*, 196-209. Retrieved from http://www.behavior-analyst-today.net/index.html

Van den Noortgate, W., & Onghena, P. (2008). A multilevel meta-analysis of single-subject experimental design studies. *Evidence-based Communication Assessment and Intervention, 2*, 142-151. doi:10.1080/17489530802505362

Van Laarhoven, T., Kraus, E., Karpman, K., Nizzi, R., & Valentino, J. (2010). A comparison of picture and video prompts to teach daily living skills to individuals with autism. *Focus on Autism and Other Developmental Disabilities, 25*, 195-208. doi:10.1177/1088357610380412

Vilardaga, R., Hayes, S. C., Levin, M. E., & Muto, T. (2009). Creating a strategy for progress: A Contextual Behavioral Science approach. *Behavior Analyst, 32*, 105-133.

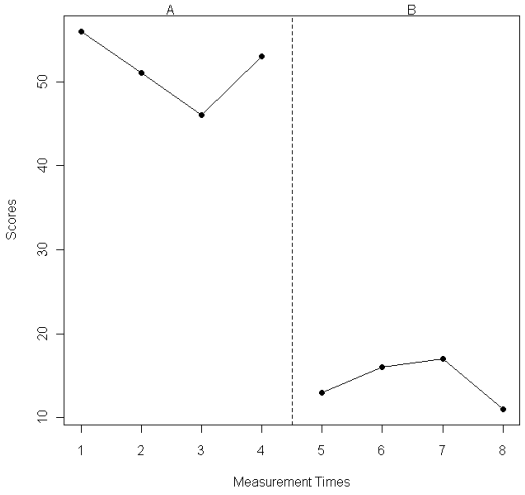
Wampold, B., & Worsham, N. (1986). Randomization tests for multiple-baseline designs. *Behavioral Assessment, 8*, 135-143.

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.

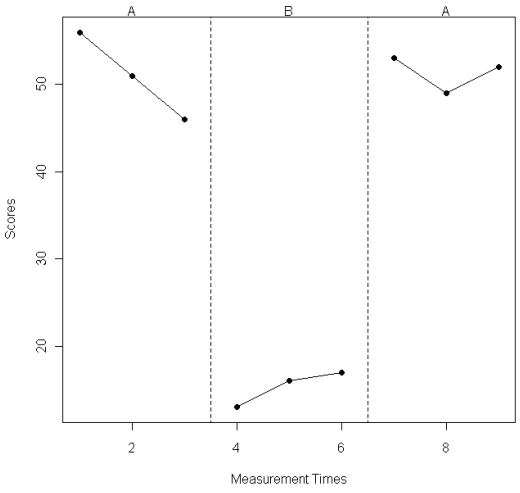
Wolery, M., Busick, M., Reichow, B., & Barton, E. E. (2010). Comparison of overlap methods for quantitatively synthesizing single-subject data. *Journal of Special Education, 44*, 18-29. doi:10.1177/0022466908328009

Zhan, S., & Ottenbacher, K. J. (2001). Single subject research designs for disability research. *Disability and Rehabilitation, 23*, 1-8.

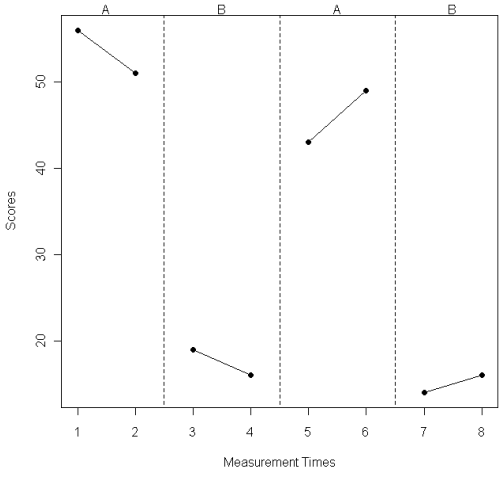
Figure 1. Examples of three common single-case phase designs: (1) AB design, i.e. the most basic phase design; (2) ABA design, i.e. reversal and withdrawal design; and (3) ABAB design. All three examples concern research on interventions for reducing challenging behavior among persons with autism. The A condition represents the control condition or baseline. The B condition represents the behavioral intervention.



(1)



(2)



(3)

Figure 2. Data set from Lorimer et al. (2002) showing the frequency of interrupting vocalizations (*y*-axis) measured on 24 days (*x*-axis) during two baseline phases (A) and two social stories intervention phases (B).

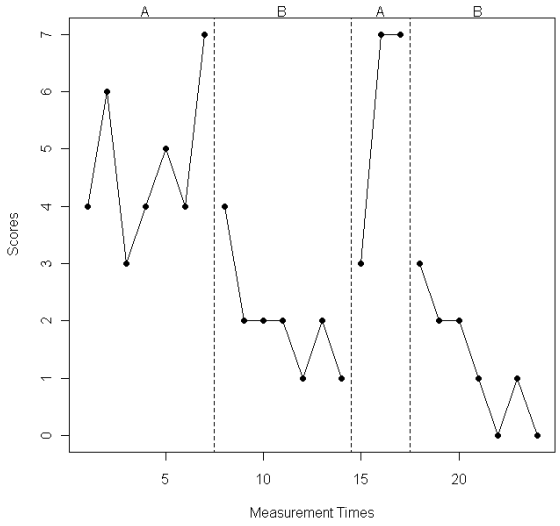


Figure 3. Example of a single-case alternation design on the effect of a behavioral intervention for reducing the challenging behavior of a person with autism. We compare the effects of the behavioral intervention condition (black dots) to the control condition (white dots).

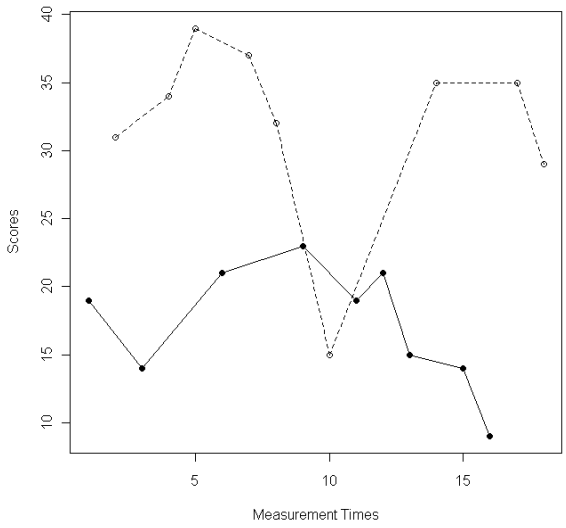


Figure 4. Data set from Van Laarhoven et al. (2010) showing the percentage of independent correct responses (*y*-axis) during the ten instructional sessions (*x*-axis) for participant Marvin. This study compared the effects of video prompting (black dots) to picture prompting (white dots).

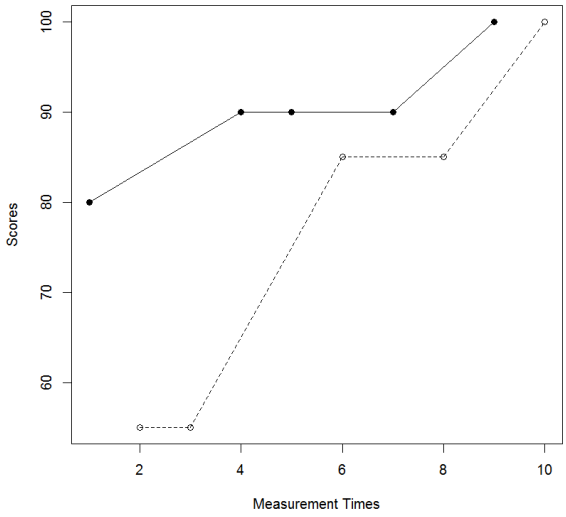


Figure 5. Visual plot of the randomization distribution for the data set from Van Laarhoven et al. (2010)

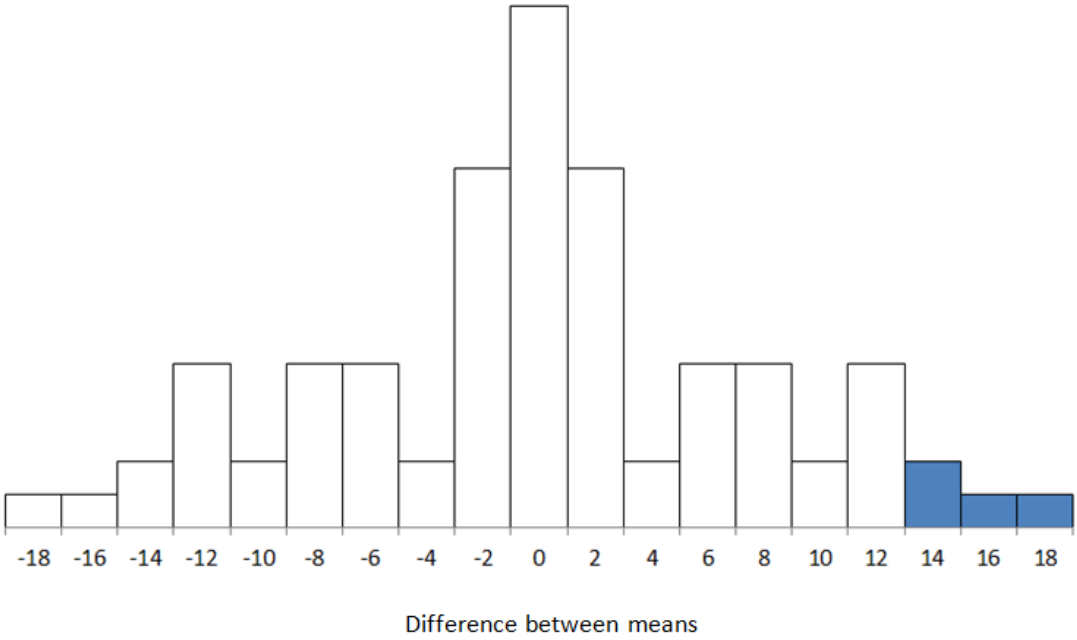


Figure 6. Example of a simultaneous replication design on the effect of a behavioral intervention for reducing the challenging behavior of four persons with autism.

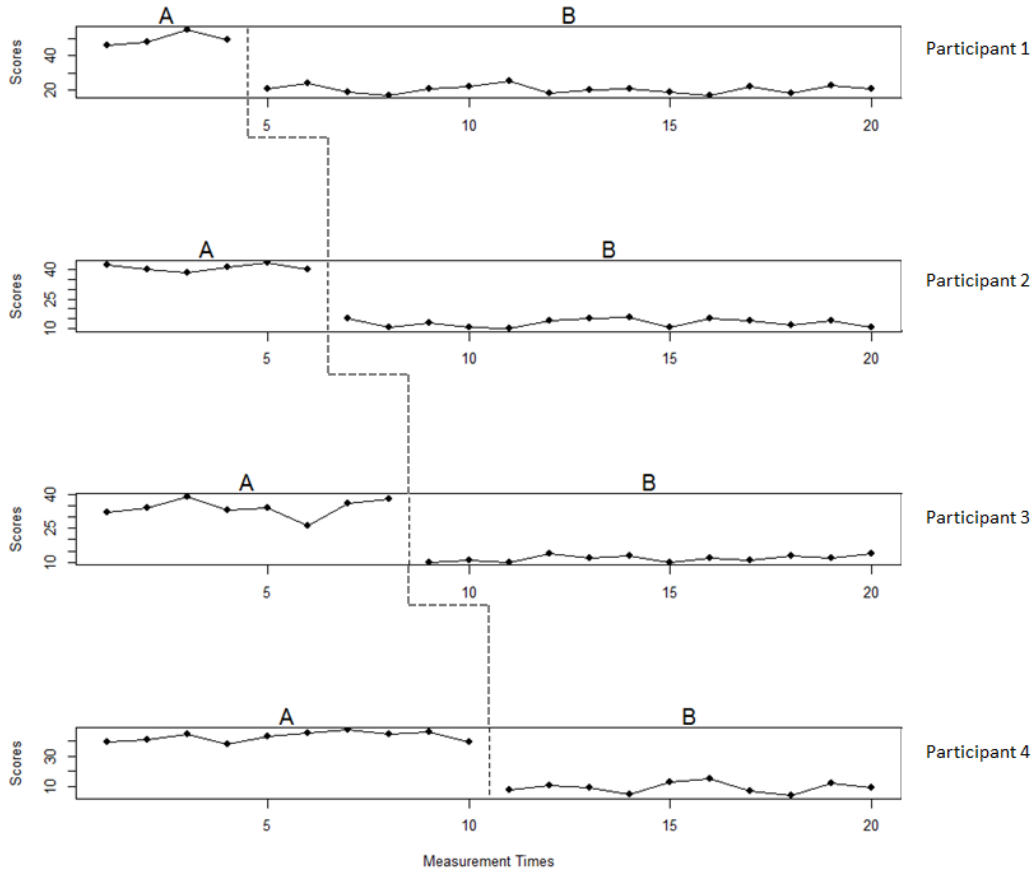
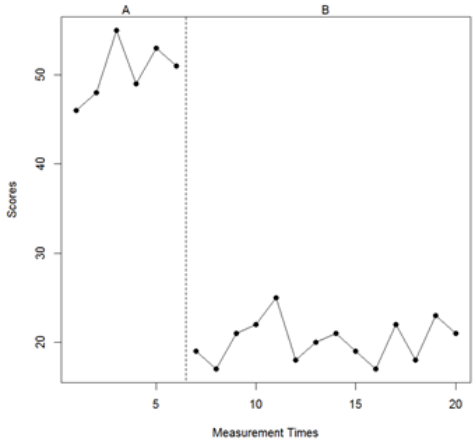
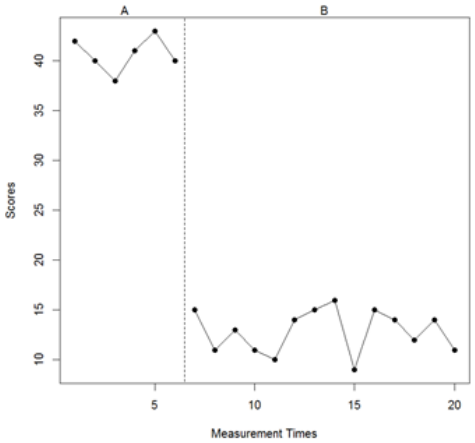


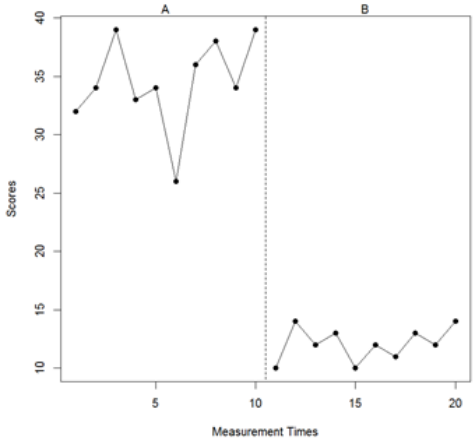
Figure 7. Example of a sequential replication design on the effect of a behavioral intervention for reducing the challenging behavior of four persons with autism.



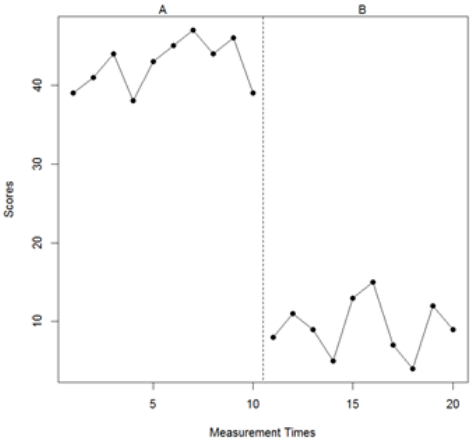
Participant 1



Participant 2



Participant 3



Participant 4

Table 1

Data Set From Lorimer et al. (2002) Collected in an ABAB Design With 24 Measurement Times

Condition	A1	A1	A1	A1	A1	A1	A1	B1	B1	B1	B1	B1	B1	B1
Score	4	6	3	4	5	4	7	4	2	2	2	1	2	1

Condition	A2	A2	A2	B2	B2	B2	B2	B2	B2	B2
Score	3	7	7	3	2	2	1	0	1	0

Table 2

Data Set From Van Laarhoven et al. (2010) for Participant Marvin Collected in an Alternating Treatments Design With Ten Instructional Sessions. This Study Compared the Effects of Video Prompting (Condition B) to Picture Prompting (Condition A)

Condition	B	A	A	B	B	A	B	A	B	A
Score	80	55	55	90	90	85	90	85	100	100

Appendix 1

R Code for the Example of Lorimer et al. (2002)

```
ABAB <- read.table(file.choose(), header = FALSE)
 library(scrt)
 quantity(design = "ABAB", MT = 24, limit = 3)
 graph(design = "ABAB", data = ABAB)
 observed(design = "ABAB", statistic = "AA-BB", data = ABAB)
 distribution.systematic(design = "ABAB", statistic = "AA-BB", limit = 3, save = "no", data =
ABAB)
 pvalue.systematic(design = "ABAB", statistic = "AA-BB", limit = 3, data = ABAB)
```

Appendix 2

R Code for the Example of Van Laarhoven et al. (2010)

```
ATD <- read.table(file.choose(), header = FALSE)
 library(scrt)
 quantity(design = "ATD", MT = 10, limit = 2)
 graph(design = " ATD ", data = ATD)
 observed(design = "ATD ", statistic = "B-A", data = ATD)
 distribution.systematic(design = "ATD ", statistic = "B-A", limit = 2, save = "no", data =
ATD)
 pvalue.systematic(design = "ATD ", statistic = "B-A", limit = 2, data = ATD)
```