Please cite as: Tanious, R., Manolov, R., & Onghena, P. (2019). The assessment of consistency in single-case experiments: Beyond A-B-A-B designs. Advance online publication, *Behavior Modification.* doi: 10.1177/0145445519882889

The assessment of consistency in single-case experiments: Beyond A-B-A-B designs

René Tanious*[1], Rumen Manolov[2], & Patrick Onghena[1]

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

[2] Department of Social Psychology and Quantitative Psychology, Faculty of Psychology, University of Barcelona

***Corresponding Author:**
René Tanious
Faculty of Psychology and Educational Sciences, KU Leuven
Tiensestraat 102 box 3762, 3000 Leuven, Belgium.
Phone: +32 16 32 82 19
E-mail: rene.tanious@kuleuven.be

Rumen Manolov
Department de Psicologia Social i Psicologia Quantitativa, Facultat de Psicologia, Universitat de Barcelona
Passaeig de la Vall d'Hebron 171, 08035, Barcelona, Spain
E-mail: rrumenovl3@ub.edu

Patrick Onghena
Faculty of Psychology and Educational Sciences, KU Leuven
Tiensestraat 102, B-3000 Leuven, Belgium
Phone: +32 16 32 59 54
E-mail: patrick.onghena@kuleuven.be

Abstract

Quality standards for single-case experimental designs (SCEDs) recommend inspecting six data aspects: level, trend, variability, overlap, immediacy, and consistency of data patterns. The data aspect consistency has long been neglected by visual and statistical analysts of SCEDs despite its importance for inferring a causal relationship. However, recently a first quantification has been proposed in the context of A-B-A-B designs, called CONsistency of DAta Patterns (CONDAP). In the current paper, we extend the existing CONDAP measure for assessing consistency in designs with more than two successive A-B elements (e.g., A-B-A-B-A-B), multiple baseline designs, and changing criterion designs. We illustrate each quantification with published research.

Keywords: Single-case experimental designs, effect sizes, consistency, statistical analysis, visual analysis, multiple baseline design, changing criterion design

Introduction

Quality standards for the conduct and analysis of single-case experimental designs (SCEDs) recommend inspecting six data aspects: level, trend, variability, overlap, immediacy, and consistency of data patterns between similar phases (Kratochwill et al., 2010). The two basic approaches for analyzing these data aspects are visual and statistical analyses, which are complementary rather than mutually exclusive (Tate et al., 2016b). Traditionally, visual analysis has been the dominant mode of analyses. However, due to the time-series nature of SCED data and varying degrees of autocorrelation visual analysis is not recommended as a stand-alone analytical methods for SCEDs (Robey, Schultz, Crawford, & Sinner, 1999). In their comparison of three existing tools for evaluating the quality and rigor of SCEDs, Zimmerman et al. (2018, p.30) found that "although each tool incorporates components of visual analysis to evaluate outcomes, quantitative measures are also frequently recommended in addition to visual analysis". Consequently, numerous quantifications have been developed over the past decades in an attempt to supplement visual analysis of each data aspect with statistical analyses. In this paper, we focus on the data aspect consistency. We present quantifications for assessing the degree of consistency in SCEDs beyond A-B-A-B withdrawal designs (cf. classification by Gast, Ledford, & Severini, 2018) and discuss the option to integrate the obtained quantifications as test statistics in randomization tests.

First, we define the data aspect consistency in its historical context and review its role in inferring a causal relationship from SCEDs. Next, we review the existing proposal for assessing consistency in A-B-A-B phase designs. We then present novel quantifications for assessing consistency in designs with more than two successive A-B elements (e.g., A-B-A-B-A-B), multiple baseline designs, and changing criterion designs.

Consistency in SCEDs: Definition, History, and Relevance

Following the widely accepted What Works Clearinghouse guidelines for the conduct and analysis of SCEDs

> 'Consistency of data in similar phases' involves looking at data from all phases within the same condition (e.g., all 'baseline' phases; all 'peer-tutoring' phases) and examining the extent to which there is consistency in the data patterns from phases with the same conditions. The greater the consistency, the more likely the data represent a causal relation (Kratochwill et al., 2010, p. 18).

Contrary to the other five data aspects, consistency between similar phases is thus assessed comparing data patterns from phases implementing the same experimental conditions. While the definition by Kratochwill et al. can probably be traced back to the guidelines for visual analysis of SCEDs offered by Horner et al. (2005), the origins of consistency as a separate data aspect date back to Baer (1977) and Parsonson and Baer (1978) who stress the importance of assessing the congruity of experimentally similar phases. It was around this time – in the 1970s and 1980s – that researchers started compiling comprehensive overviews of SCE designs and analysis (Onghena, Tanious, De, & Michiels, 2019). Consistency is thus one of the oldest data aspects recommended for analysis of SCEDs. Furthermore, as the definition by Kratochwill et al. highlights, greater consistency between experimentally similar phases is an important moderator of a causal relationship between the manipulation of the independent variable and the scores on the dependent variable.

A few years after Parsonson and Baer, Kazdin (1982) was among the first researchers to discuss a second type of consistency in SCED data: the consistency of the effects. Contrary to the consistency between experimentally similar phases, consistency of the effects is assessed for each potential demonstration of an effect:

> Consistency refers to […] the extent to which changes (in level, trend, or variability) are the same for each potential demonstration of effect. In SCD research, the critical factor in determining a functional relation is the consistency of behaviour change between conditions; consistent but small changes in level between conditions are superior to inconsistent changes of larger magnitude. (Ledford, Lane, & Severini, 2018, pp. 6-7)

This second type of consistency can increase our confidence in the existence of a causal relationship just as much as the consistency across experimentally similar conditions. As Ledford et al. (2018) highlight, consistent but small changes are more indicative of a successful intervention than large changes than cannot be replicated.

Distinguishing Two Types of Consistency for Different SCE Designs

It is important to clearly distinguish between the two types of consistency. The consistency of data patterns is not necessarily an aim of an experimental intervention. It is rather an indication that the introduction of the intervention (and possibly the withdrawal depending on the design) led to a similar pattern of responding. It should be highlighted that a similar pattern of responding is conceptually different from low variability. It can, for example, happen that a subject has highly variable scores on a dependent variable in the first intervention phase (e.g., 3,15,8,21). If these scores are replicated in the second intervention phase, the consistency of data patterns across phases representing the same condition is very high (indicative of a similar pattern of responding) even if the variability in each individual phase is high as well. Variability can be assessed within phases and changes in variability can be compared across phases, whereas the consistency of data patterns is, per definition, assessed between phases. Contrary to the consistency of data patterns, the consistency of the effects is a direct result of the intervention. For example, the intervention could lead to a consistent change in mean scores between adjacent phases each time the intervention is

introduced. Moreover, note that a difference in variability (e.g., a reduction) across phases belonging to different conditions could be an aim for the intervention, whereas the consistency of data patterns from the same condition is not.

Furthermore, it is noteworthy that both types of consistency are conceptually different for different SCE designs. While in phase designs and changing criterion designs intervention effectiveness is demonstrated within participants, in multiple baseline designs intervention effectiveness is demonstrated both within and between participants, behaviors, or settings, which Hayes (1981) refers to as the combined-series strategy. For multiple baseline designs, the consistency of data patterns is compared between baseline phases and between intervention phases for different participants, behaviors, or settings. If in a multiple baseline across participants design the participants differ in their initial baseline measures, then the consistency of experimentally similar phases will be low as well. In a situation like that, the lack of consistency does not necessarily entail a lack of experimental control. Given that all participants started with a different baseline level, it might therefore be expected that the consistency of the data patterns remains low for the experimental phases. Similarly, in multiple baseline design it is less likely that the intervention produces a consistent effect for all participants, behaviors, or settings. However, if a consistent effect is produced for participants starting with different initial baseline levels, this would entail different intervention phase levels (i.e., lack of consistency of data patterns within the intervention condition). In phase designs and changing criterion designs however, where intervention effectiveness is demonstrated within participants, a lack of consistency of data patterns can be more indicative of a lack of experimental control. The exception is phase designs in which a return to initial baseline levels is not possible during subsequent baseline phases (e.g., because the change in the dependent variable is irreversible). A final remark on the distinction between the two types of consistency concerns only the changing criterion design. It might be

argued that, by definition, the two types of consistency are conceptually the same for the changing criterion design. The consistency of data patterns in the changing criterion design can be understood in terms of the extent to which the measurements align consistently with the given criterion in each phase. For the changing criterion design, this is conceptually the same as the consistency of the effect. The present paper explicitly deals with methods for assessing the consistency of data patterns only. In the next section, we present an existing proposal for assessing the consistency of data patterns in A-B-A-B designs.

Analyzing the Consistency of Data Patterns in SCEDs

Historically, visual analysis has been the dominant mode of analysis for SCEDs. Probably the earliest systematic review of SCED analytical techniques employed by applied researchers was conducted by Kratochwill and Brody (1978). Kratochwill and Brody reviewed four leading behavior modification journals for SCED studies published between 1963 and 1974 and found that the proportion of studies employing statistical inference ranged from 4% to 9%. Given these findings, Kratochwill and Brody concluded that “there is a need for more attention to the technology of graphing. In a science that depends so heavily on visual analysis […] and its related statistical properties, it is necessary to derive some major guidelines in this area” (p. 302). At the same time, Kratochwill and Brody issued a warning that sole reliance on visual analysis may lead to erroneous conclusions as evidence about the flaws of visual analysis and need for statistical inference criteria started accumulating in the 1970’s (DeProspero & Cohen, 1979; Gentile, Roden, & Klein, 1972; Jones, Vaught, & Weinrott, 1977; Jones, Weinrott, & Vaught, 1978). In this context, the earliest systematic guidelines for visual analysis of SCEDs were proposed by Parsonson and Baer (1978), which included, as previously mentioned, the first type of the data aspect consistency. It is this type of consistecy, the consistency of data patterns in experimentally similar phases, that we focus on

in the present article. For existing proposals on the consistency of the effects, the interested reader is referred to Manolov (2018) for multiple baseline designs and Tanious, De, Michiels, Van den Noortgate, and Onghena (2019a) for A-B-A-B phase designs.

Regarding the consistency of data patterns in experimentally similar phases, only one quantification has been proposed so far. Tanious et al. (2019) developed the CONsistency of DAta Patterns (CONDAP) measure for quantifying the degree of consistency between data patterns in A-B-A-B phase designs based on the Manhattan distance. The basic premise of CONDAP is that if two data patterns are highly consistent, then the standardized average Manhattan distance should be accordingly low. Therefore, lower CONDAP values indicate higher consistency and higher CONDAP values indicate lower consistency. For A-B-A-B phase designs, CONDAP can be calculated as shown in Equation 1.

$$CONDAP = \frac{\frac{1}{k\,n_s}\sum_{j=1}^{k}\sum_{i=1}^{n_s} \mid s_{ij} - l_{ij} \mid}{\sqrt{\frac{(n_s-1)*SD_s^2 + (n_l-1)*SD_l^2}{n_s+n_l-2}}} \quad (1)$$

The numerator in Equation 1 is the mean Manhattan distance across all comparisons of k sequences of equal length n for the two phases being compared. In an A-B-A-B design, for each condition, there are two experimentally similar phases. If both phases have the same number of data points, the number of compared sequences is always equal to one and only one comparison of the two phases is necessary. If the two phases differ in lengths, the longer phase is shortened to a sequence that is of equal lengths to the shorter phase. The shorter phase is then compared to each sequence of the longer phase that is equal to the lengths of the shorter phase. $S_{ij}$is the *i*th data point for the *j*th sequence from the shorter phase and l_{ij} is the *i*th data point for the *j*th sequence from the longer phase. Figure 1 visualizes the calculation of CONDAP for an A-B-A-B design where the B1 phase has five measurements (8,6,7,7,6) and

the B2 phase has three measurements (7,5,9). The CONDAP for Figure 1 can then be calculated as follows. First, the sum of the vertical dashed lines (the Manhattan distances) for each panel is calculated and divided by three (which is the length of the shorter phase) to obtain the average Manhattan distance per comparison of equal length of the two data patterns. Between each panel, the shorter phase is shifted by one measurement occasion to the right. Second, the three average Manhattan distances are then summed up and divided by the number of comparisons, in this case three. Finally, this overall average Manhattan distance is then divided by the pooled standard deviation of the two phases as proposed by Van den Noortgate and Onghena (2008) to obtain the scale invariant CONDAP (whose value would be 1.16 for the current example). If all experimentally similar phases have a standard deviation of zero, then the denominator of CONDAP would be zero as well, and it is recommended to calculate the overall average Manhattan distance without standardization (Tanious et al., 2019)

[INSERT FIGURE 1 HERE]

Based on a systematic review of a sample of 119 applied A-B-A-B studies, Tanious, De, Michiels, Van den Noortgate, and Onghena (2019b) proposed the following guidelines for interpreting CONDAP: very high, $0 \leq CONDAP \leq 0.5$; high, $0.5 < CONDAP \leq 1$; medium, $1 < CONDAP < 1.5$; low, $1.5 < CONDAP \leq 2$; very low, $CONDAP > 2$. The two data patterns in Figure 1 are thus medium consistent. It is important to note that these guidelines were developed specifically for designs using successive A-B comparisons. We will use these guidelines throughout the article for other designs for illustrative purposes and comparability, but they should be interpreted with caution.

Consistency of Data Patterns: When Should it Be Assessed?

It is noteworthy that the What Works Clearinghouse panel uses an A-B-A-B phase design as an example for demonstrating their developed guidelines. The immediate question following

from that is whether consistency is only desirable in A-B-A-B designs. The definition by Kratochwill et al. refers to ‘consistency of data in similar phases’. From this statement, it can be deduced that each manipulation of the independent variable must occur at least twice in order to assess consistency. By definition, this excludes phase designs such as A-B, A-B-C, A-B-C-D, and A-B-BC-C in which each unique manipulation of the independent variable is only introduced once. In addition, there are phase designs in which only one of the manipulations of the independent variable is introduced twice. Examples of such designs are A-B-A, B-A-B, A-B-C-A, and A-B-C-B. In such designs, the consistency of data patterns can only be selectively assessed for the experimental condition that was introduced twice. Furthermore, even in phase designs where each manipulation of the independent variable occurs at least twice (e.g., A-B-A-B), consistency cannot always be assessed, in relation to the number of measurements available. In line with quality standards for SCEDs, each phase should contain a minimum of three data points to meet minimum evidence standards (e.g., Beeson & Robey, 2006; Kratochwill, et al., 2010) and five data points per phase are required to meet evidence standards without reservation (Ganz & Ayres, 2018; Horner et al., 2005; Kratochwill et al., 2010, 2013; Tate et al., 2016b; U.S. Department of Education, 2016). Additionally, consistency is assessed by comparing data patterns that emerge over time making it implausible to quantify consistency for very short time-series. For these reasons, we recommend against quantifying consistency of data patterns with less than three data points per phase. By the same logic, it is not plausible to assess the consistency of data patterns for one particular kind of SCEDs: the alternating treatments designs. This design utilizes rapid and repeated manipulations of two or more conditions (Wolery, Gast, & Ledford, 2018) and “the crucial factor of the design is the unique intervention phase in which two separate interventions are administered concurrently” (Kazdin, 2011, p. 198). Due to the rapid alternation of treatments, there are no distinguishable phases in this design so that no

consistent data patterns across experimentally similar phases can emerge. In the following sections, we extend the existing CONDAP measure for designs in which consistency of data patterns is desirable: multiple baseline designs, phase designs with more than two consecutive A-B comparisons, and changing criterion designs.

Extensions of CONDAP for more than four Phases and Multiple Baseline Designs

The basic A-B-A-B phase design can be extended by as many phases as the researcher believes are necessary and feasible. One such extension is the A-B-A-B-A-B design in which one more baseline and one more experimental phase are added to the A-B-A-B structure. The advantages of adding one more A-B pattern include the following: repeated possibilities for demonstrating experimental control over the dependent variable, extended study until full clinical treatment has been achieved, and added flexibility (Barlow, Nock, & Hersen, 2009). Figure 2 shows an example of an applied study from the autism literature using this design. Angell, Nicholson, Watts, and Blum (2011) examined the effectiveness of a multicomponent adapted Power Card strategy to decrease latency during interactivity transitions for three children with developmental disabilities. The dependent variable was interactivity transitions that occurred within the classroom. This was measured as the latency of students with developmental disabilities in response to teacher cues to initiate classroom interactivity transitions. The baseline phases included typical classroom conditions. During intervention phases, several components were added to the students' routines which included presentation of the Power Card, verbal cues by the teacher, and verbal praise by the teacher. Figure 2 shows the data of Quincy, an 11-year-old male diagnosed with autism.

[INSERT FIGURE 2 HERE]

As the definition by Kratochwill et al. (2010) suggests, consistency of similar data patterns is assessed separately for each manipulation of the independent variable. In the example in Figure 2, we can thus assess the consistency of data patterns for the three baseline phases and

the three intervention phases separately. This can be achieved by adding one additional step to the CONDAP formula for A-B-A-B designs. Whereas in an A-B-A-B design only one comparison is needed between the $A_1$and A_2 phase on the one hand and the B_1 and B_2 phase on the other, an A-B-A-B-A-B requires additional comparisons to incorporate the A_3 and B_3 phases. To achieve this, we can first calculate separate CONDAP values for each comparison between A_1 and A_2, between A_1 and A_3, and between A_2 and A_3. The consistency matrix below shows the results of this first step.

	A_1	A_2	A_3
A_1	0	1.06	2.37
A_2		0	1.65
A_3			0

The diagonal values of the matrix equal zero because each phase is perfectly consistent with itself (i.e., identical data patterns) as indicated by a CONDAP value of zero. Subsequently, an average across all three comparisons can be calculated to obtain the overall consistency of baseline data patterns: (1.06 + 2.37 + 1.65) / 3 = 1.69. The baseline data patterns thus show low consistency using the interpretative guidelines developed for A-B-A-B designs. Following the same steps for the B-phases leads to obtaining a CONDAP value of 1.45 indicative of medium consistency. Beyond the application of the interpretative benchmarks, the consistency is lower for the baseline data patterns and this agrees well with the visual analysis of the data. Visual inspection of Figure 2 reveals that the A-phases are all highly variable and show differences in level whereas the B-phases are less variable and differences in level are smaller than between A-phases. The calculated CONDAP values express these differences in terms of higher consistency between experimental phases than baseline phases.

For assessing the consistency in multiple baseline designs, the same steps can be followed. Multiple baseline designs across subjects can be defined as follows.

> In the multiple baseline design across subjects, each individual targeted for treatment is exposed to the same environment. Treatment is delayed for each successive subject in time-lagged fashion because of the increased length of baselines required for each. The functional relationship between treatment and target behavior can be determined only when such treatment is applied to each subject in succession. (Barlow et al., 2009, p. 234)

If the goal is to change several distinct target behaviors in a single participant, a multiple baseline across behaviors design can be used (Rvachew, 1988). In such a design, the intervention is introduced in a time-lagged fashion to different target behaviors to demonstrate experimental control of the intervention over each target behavior. Figure 3 shows an example of a multiple baseline across subjects design. Scheeler, Morano, and Lee (2018) used this design to investigate the effectiveness of a training program for four paraeducators working with students diagnosed with Autism spectrum disorder. During baseline sessions, the paraeducators received only delayed feedback from a special education classroom teacher. During intervention sessions, the paraeducators received immediate feedback from the special education classroom teacher via bug-in-ear technology. The dependent variable displayed in Figure 3 is percentage of contingent specific praise statements delivered by the paraeducator to the students reported as the percentage of totals praise statements.

[INSERT FIGURE 3 HERE]

To calculate the consistency across data patterns of experimentally similar phases, we can follow the same logic as in the A-B-A-B-A-B example. The only difference is that in a multiple baseline across subjects design, we compare experimentally similar phases across subjects instead of within a subject. Given that Scheeler et al. (2018) worked with four participants, we need to add one more row and column to the consistency matrix for A-B-A-B-A-B designs to incorporate the additional comparisons. This also holds true for designs

consisting of adjacent A-B comparisons with more than six phases. The consistency matrix below shows the results of all consistency comparisons for the intervention phases of the Scheeler et al. data.

	B_1	B_2	B_3	B_4
B_1	0	1.33	1.02	1.28
B_2		0	1.15	1.67
B_3			0	1.20
B_4				0

In a second step, which is identical to the consistency assessment for A-B-A-B-A-B designs, we can calculate the average across all comparisons: (1.33 + 1.02 + 1.28 + 1.15 + 1.67 + 1.20) / 6 = 1.27. The consistency is higher than, for example, in the data patterns of the Angell et al. data set. Following the same steps for the baseline phases, leads to obtaining a CONDAP value of 1.17. In general, the number of consistency comparisons *c* per experimentally similar phase –the upper right triangle in the consistency matrix– for A-B-A-B-A-B and multiple baseline designs equals:

$$c = \frac{\left(n_p^2 - n_p\right)}{2} \tag{2}$$

In Equation 2, n_p represents the number of experimentally similar phases. For example, the Scheeler et al. dataset contained four phases for each intervention. The number of consistency comparisons *c* for each experimentally similar phase thus equals $\frac{\left(4^2-4\right)}{2}=6.$

Extensions of CONDAP for Changing Criterion Designs

The changing criterion design has been introduced as a distinct design in a seminal paper by Hartmann and Hall (1976). As the name suggests, in this design, an individual is subjected to

changing criterions (goals) for the rate of the target behavior. Barker, McCarthy, Jones, and Moran (2011) describe the procedures for this design as follows:

> A criterion is set that represents a target (goal) for the participant to meet. This criterion (or goal) will change throughout the course of the study. It is anticipated that the variable improves in increments to match the criterion that is specified as part of the intervention (Kazdin, 1982). Normally, rewards or incentives are provided to facilitate the attainment of a designated criterion […] In the changing criterion design, the required level of a target variable is altered repeatedly (e.g., increasing the amount of daily exercise time) to improve performance of this variable over time. (p. 109)

The changing criterion design is especially useful when immediate large changes in target behavior are either impossible or undesirable because in this design the researcher can apply gradual shifts toward a desired goal (Klein, Houlihan, Vincent, & Panahon, 2017). When the dependent variable in a changing criterion design changes according to the predetermined criterion levels, an intervention effect is demonstrated (Kinugasa, Cerin, & Hopper, 2004). Figure 4 shows an example of an applied changing criterion design. Voulgarakis and Forte (2015) used this design to examine the effectiveness of an escape extinction and negative reinforcement-based approach to treating food refusal in a child with cerebral palsy. The dependent variable displayed in Figure 4 is the number of instances of the child depositing a bite of food into his mouth during a 30-minute period. During the first phase, baseline measures were collected with no intervention taking place. The criterions for the subsequent intervention phases were 5, 7, 10, 7, and 12 bites. During intervention phases, the child received positive verbal reinforcement and was allowed to exit the meal as well as meal area upon reaching the criterion. The upper left panel of Figure 4 shows the raw data.

[INSERT FIGURE 4 HERE]

In such a design, the assessment of consistency is highly desirable. If the rates of target behavior displayed by the participant are *consistent* with the criteria set by the researcher, then our confidence in the existence of a causal relationship increases.

However, since the criterion changes in each consecutive phase, the assessment of consistency for this design differs from the assessment of consistency in the previously discussed phase designs and multiple baseline designs for two reasons. First, there is only one baseline phase in the beginning of the study in which no intervention is present. After the initial baseline phase, no withdrawal of the intervention takes place (Barlow et al., 2009). The assessment of consistency of baseline phases is thus neither possible nor desirable for the changing criterion design. Second, in a changing criterion design the criterion rate for the target behavior set by the researcher changes for each consecutive phase. Therefore, the data patterns themselves are expected to be *in*consistent if the intervention is successful and the rate of target behavior changes with each criterion change. Accordingly, we recommend inspecting the consistency of data patterns in relation to the criterion in each phase rather than the consistency of the raw data. Moreover, unlike A-B-A-B, A-B-A-B-A-B, and multiple-baseline designs, the assessment of consistency in a changing criterion design would not require distinguishing between consistency of data in similar phases (which is the focus of the current text) and consistency of effects. Actually, for a changing criterion design, there would be only one kind of consistency: the degree to which the measurements obtained match the criterion levels set by the researcher, in the different subphases of the intervention phase. In that sense, the assessment of consistency in a changing criterion design, as described here, is even more important for inferring a causal relation between the reinforcer and the target behavior, because it does not need to be complemented with a second type of consistency.

One measure of intervention effectiveness for changing criterion designs that is sensitive to the differences between the scores and the criterion is the Mean Absolute

Difference which “is calculated by taking the difference between the scores and the criterion at each measurement occasion, dropping the sign, summing all these absolute differences, and dividing by the total number of scores” (Onghena et al., 2019, p. 4). Following the logic of the Mean Absolute Deviation, we propose to use the difference between the scores and the criterion in each phase for assessing the consistency of data patterns with a slight modification. Instead of taking the absolute difference, we take the difference between each score and the criterion *as is* in order to preserve the original data patterns. The added vertical dashed lines in upper right panel of Figure 4 show these differences. For example, for the first experimental subphase the difference between the scores and the criterion of five are 0, 1, and 0. Converting each score to a difference score in this way gives the data shown in the lower left panel of Figure 4 whereby the baseline phase is dropped because no intervention or criterion was present in that phase and not assessment of consistency is possible. Subsequently, the difference scores can be used to assess the consistency between all pairs of subphases in the same way as for A-B-A-B-A-B and multiple baseline designs. For the scores depicted in Figure 6 we obtain the following consistency matrix where the numbers indicate the order in which the subphases occurred during the experiment.

	1	**2**	**3**	**4**	**5**
1	0	1.15	0.58	0	1.63
2		0	1.73	1.15	1.63
3			0	0.58	1.22
4				0	1.63
5					0

Taking the average across all comparisons gives a CONDAP of 1.13 indicative of medium consistency when applying the benchmarks developed for A-B-A-B designs.

In general, the number of consistency comparisons c for changing criterion designs equals:

$$c = \frac{\left(n_p - 1\right)^2 - n_p}{2} \tag{3}$$

For changing criterions designs, the baseline phases needs to be subtracted from the overall number of phases before squaring the number of phases.

Discussion

Historically, the consistency of data patterns across experimentally similar phases has a long tradition in the visual analysis of SCED data. In line with recent literature acknowledging the complementarity of visual and statistical analyses, the aim of the present paper was to extend the existing CONDAP measure for quantifying consistency in A-B-A-B designs to designs with more than two adjacent AB comparisons, multiple baseline designs, and changing criterion designs. Using published data sets for illustrative purposes, we showed how in each design consistency can be assessed by means of a consistency matrix. Moreover, we have developed R code for obtaining the quantifications: A user-friendly web-application implementing CONDAP is freely available at https://manolov.shinyapps.io/CONDAP/. The user has to specify the data file as indicated in the examples and, the software provides the CONDAP matrix and the average CONDAP values. The web application further provides graphical representations of baseline phases' consistency and intervention phases' consistency. In addition, the user can generate a time series graph of the raw data.

The focus of this article was on the data aspect consistency because this data aspect has only recently been quantified for the first time and only in the context of A-B-A-B designs. Given the importance of the data aspect consistency for inferring a causal relationship in SCEDs, extensions of the existing CONDAP for A-B-A-B designs to other designs can have great added value. Notwithstanding this importance of consistency in analyzing data obtained from SCEDs, it is important to keep in mind that consistency of data in similar phases should always be assessed alongside the other five data aspects suggested in the What Works Clearinghouse guidelines (Kratochwill et al., 2010): level, trend, variability,

overlap, and immediacy, as well as the consistency of effects. An inferential testing procedure for all data aspects has been proposed by Tanious, De, and Onghena (2019).

Furthermore, inspecting the consistency matrix itself rather than the average CONDAP only can lead to a better understanding of the effects of a new treatment and experimental control. For example, in the changing criterion design, the overall CONDAP indicated medium consistency when using the guidelines for A-B-A-B designs for illustrative purposes. Looking at the consistency matrix, we can see that intervention subphases three and four are highly consistent which is remarkable because subphase three was an increase in the criterion whereas subphase four was a decrease in the criterion. Such consistency across different criterion changes, including a reversal to a lower level, is indicative of experimental control (cf. Kinugasa et al., 2004).

It should also be noted that CONDAP can be extended to extensions of the designs presented in this article. One such extension is the range-bound changing criterion design in which a lower and upper bound are specified for each intervention phase (McDougall, 2005). McDougall suggests an upper bound of 10% above the desired mean for each phase and a lower bound 10% beneath the desired mean for each phase. According to McDougall, advantages of the range-bound changing criterion design include: establishing a ceiling for acceptable improvement to prevent counter-therapeutic effects (e.g., to avoid injury in sports training interventions), establishing a floor of acceptable performance for gradual improvement, and added flexibility to adapt to the subject's individual circumstances. CONDAP can be calculated for this design by assigning a value of zero for all data points within the acceptable range, assigning a positive value for each score equal to the distance on the y-axis above the upper limit, and assigning a negative value for each score equal to the distance on the y-axis under the lower limit. With these distance scores for the range-bound

changing criterion design, CONDAP can be calculated in the same way as for the single-point changing criterion design.

Limitations and Future Research

The CONDAP extensions presented in this article are subject to a few limitations that offer potential avenues for future research. First, each extension of CONDAP was demonstrated stepwise using only one published applied data set to explain each extension in depth. Future studies could integrate the presented CONDAP extensions in simulation studies or systematic reviews. Such studies could help in cross-validating the measures with trained visual analysts and find out what are the typical CONDAP values found in published literature. A second related limitation is that the presented guidelines for interpreting CONDAP values were developed based on a systematic review of A-B-A-B designs. We do not recommend a straightforward application of these guidelines to other designs as this might systematically over- or underestimate the degree of consistency in experimentally similar phases. As discussed previously, neither the underlying logic, nor the demonstration of an intervention effect is the same in an A-B-A-B design, multiple-baseline design, and in a changing criterion design. This is also why the CONDAP calculation for changing criterion designs is slightly different. Moreover, it has to be taken into account that a multiple-baseline design usually entails across participants replication, whereas A-B-A-B and changing criterion designs replicate within a participant. The development of interpretative CONDAP guidelines specific to each design is therefore an important challenge for future research. Again, systematic reviews and cross-validation with trained visual analysts might help in developing guidelines specific to each design. A third potential avenue for future research refers to the type of consistency being assessed. In the presented study, we focused solely on the consistency of data patterns in experimentally similar phases. A second -and arguably equally important-

type of consistency is the consistency of the effects. In the context of A-B-A-B designs, Tanious et al. (2019a) proposed the CONsistency of the EFFects (CONEFF) measure for assessing the change in each data aspect between similar phase changes. Future studies might focus on extending CONEFF to the designs presented in this article so that both types of consistency can be assessed. A fourth potential avenue for future research is the development of software to integrate the presented consistency measures as test statistics in randomization tests.

Conclusion

Consistency of experimentally similar phases is an important indicator of a causal relationship in SCEDs. In this article, we presented extensions of the existing CONDAP measure for assessing consistency in phase designs with more than two successive A-B comparisons, multiple baseline designs, and changing criterion designs. Calculating CONDAP alongside effect size measures for level, trend, variability, overlap, and immediacy can be valuable supplements to mere visual analysis. In the online supplementary material, we provide generic R-code for executing all analyses presented in this article.

References

Angell, M. E., Nicholson, J. K., Watts, E. H., & Blum, C. (2011). Using a multicomponent adapted power Card strategy to decrease latency during interactivity transitions for three children with developmental disabilities. *Focus on Autism and Other Developmental Disabilities, 26,* 206-217.

Baer, D. M. (1977). "Perhaps it would be better not to know everything". *Journal of Applied Behavior Analysis, 10,* 167-172. doi:10.1901/jaba.1977.10-167.

Barker, J., McCarthy, P., Jones, M., & Moran, A. (2011). *Single-case research methods in sport and exercise psychology.* New York, NY: Routledge.

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

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

DeProspero, A., & Cohen, S. (1979). Inconcistent visual analysis of intrasubject data. *Journal of Applied Behavior Analysis, 12*, 573-579.

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

Gast, D. L., Ledford, J. R., & Severini, K. E. (2018). Withdrawal and reversal designs. In J. R. Ledford, & D. L. Gast, *Single case research methodology: Applications in special education and behavioral sciences (3rd ed.)* (pp. 215-239). New York & Abingdon: Routledge.

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.

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

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

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. doi:10.1177/001440290507100203

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

Jones, R. R., Weinrott, M. R., & Vaught, R. S. (1978). Effects of serial dependency on the agreement between visual and statistical inference. *Journal of Applied Behavior Analysis, 11*, 277-283.

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

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

Kinugasa, T., Cerin, E., & Hopper, S. (2004). Single-subject research designs and data analyses for assessing elite athletes' conditioning. *Sports Medicine, 34*, 1035-1050.

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

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 What Works Clearinghouse: http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf

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

Ledford, J. R., Lane, J. D., & Severini, K. E. (2018). Systematic use of visual analysis for assessing outcomes in single case design studies. *Brain Impairment, 19,* 1-14. doi:10.1017/BrImp.2017.16.

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

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

Michiels, B., Heyvaert, M., Meulders, A., & Onghena, P. (2017). Confidence intervals for single-case effect size measures based on randomization test inversion. *Behavior Research Methods,49,* 363-381. doi:10.3758/s13428-016-0714-4.

Onghena, P., Tanious, R., De, T. K., & Michiels, B. (2019). Randomization tests for changing criterion designs. *Behaviour Research and Therapy, 117,* 18-27. doi:10.1016/j.brat.2019.01.005.

Parsonson, B., & Baer, D. (1978). The analysis and presentation of graphic data. In T. Kratchowill, *Single Subject Research* (pp. 101–166). New York: Academic Press.

Robey, R. R., Schultz, M. C., Crawford, A. B., & Sinner, C. A. (1999). Review: Single-subject clinical-outcome research: Designs, data, effect sizes, and analyses. *Aphasiology, 13,* 445-473.

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

Scheeler, M. C., Morano, S., & Lee, D. L. (2018). Effects of immediate feedback using bug-in-ear with paraeducators working with students with autism. *Teacher Education and Special Education, 41,* 24-38. doi:10.1177/0888406416666645.

Tanious, R., De, T. K., & Onghena, P. (2019). A multiple randomization testing procedure for level, trend, variability, overlap, immediacy, and consistency in single-case phase designs. *Behaviour Research and Therapy, 119,* Advance online publication. doi:10.1016/j.brat.2019.103414.

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

Tanious, R., De, T. K., Michiels, B., Van den Noortgate, W., & Onghena, P. (2019b). Consistency in single-case ABAB phase designs: A systematic review. *Behavior Modification,* Advance online publication. doi:10.1177/0145445519853793.

U.S. Department of Education, Institute of Education Sciences, *What Works Clearinghouse.* (2019, March). Retrieved from https://ies.ed.gov/ncee/wwc/Docs/ReferenceResources/wwc_srg_scd_instructions_s3_v2.pdf

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.

Voulgarakis, H., & Forte, S. (2015). Escape extinction and negative reinforcement in the treatment of pediatric feeding disorders: A single case analysis. *Berhavior Analysis in Practice, 8*, 212–214. doi:10.1007/s40617-015-0086-8.

Wolery, M., Gast, D. L., & Ledford, J. R. (2018). Comparitive Designs. In J. R. Ledford, & D. L. Gast, *Single case research methodology: Applications in special education and behavioral sciences* (pp. 283-334). New York & Milton Park: Routledge.

Zimmerman, K. N., Ledford, J. R., Severini, K. E., Pustejovsky, J. E., Barton, E. E., & Lloyd, B. P. (2018). Single-case synthesis tools I: Comparing tools to evaluate SCD quality and rigor. *Research in Developmental Disabilities, 79*, 19-32. doi:10.1016/j.ridd.2018.02.003.

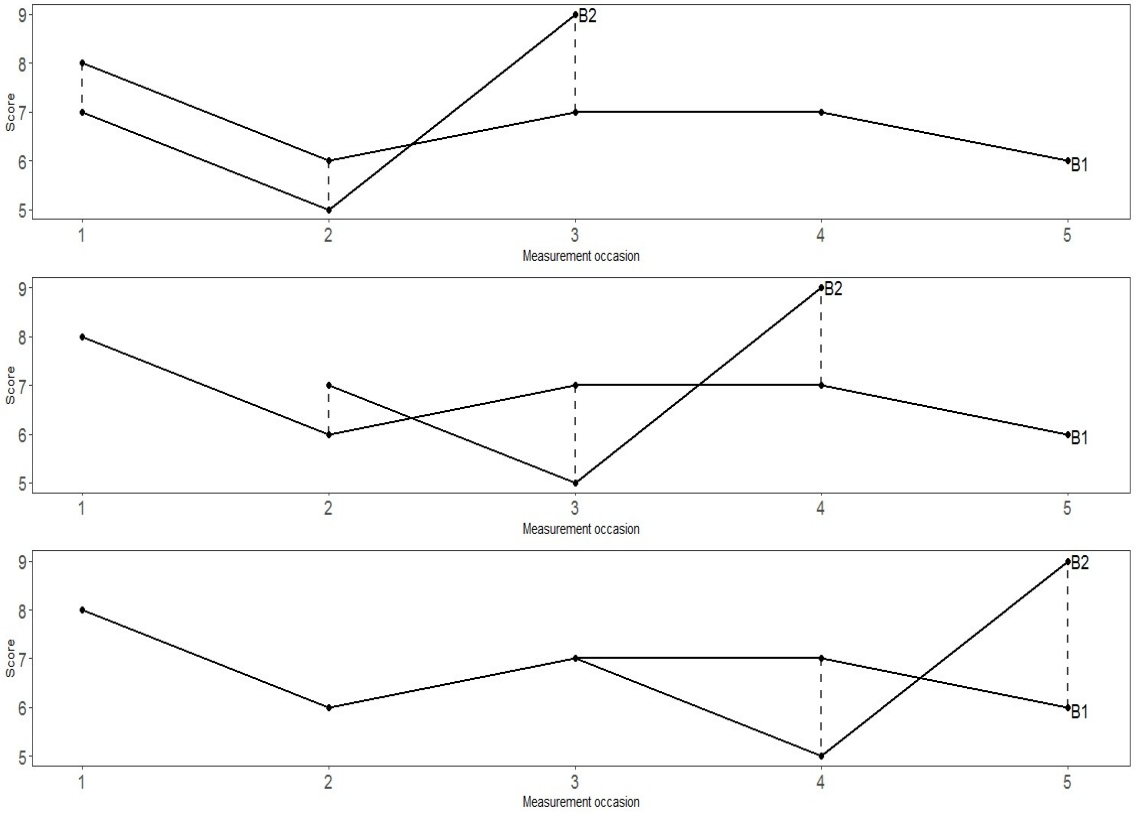

Figure 1. CONDAP calculation for an A-B-A-B phase design with unequal phase lengths

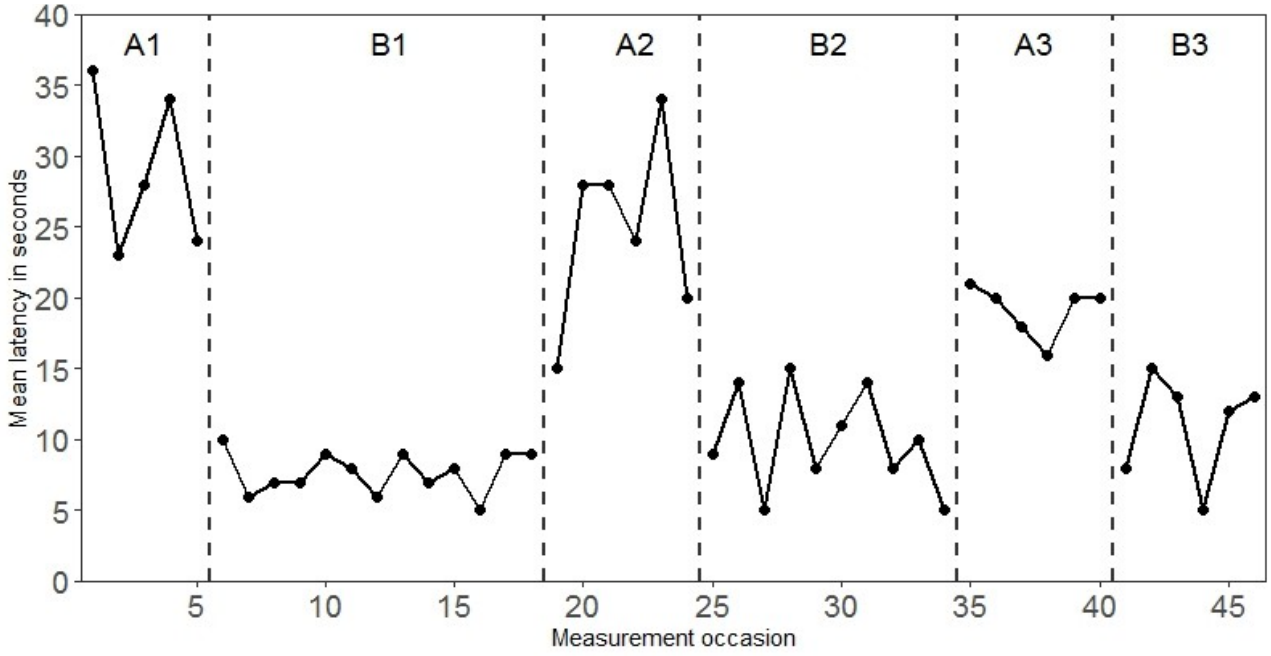

Figure 2. Example of an A-B-A-B-A-B design. Data from Angell, Nicholson, Watts, and Blum (2011)

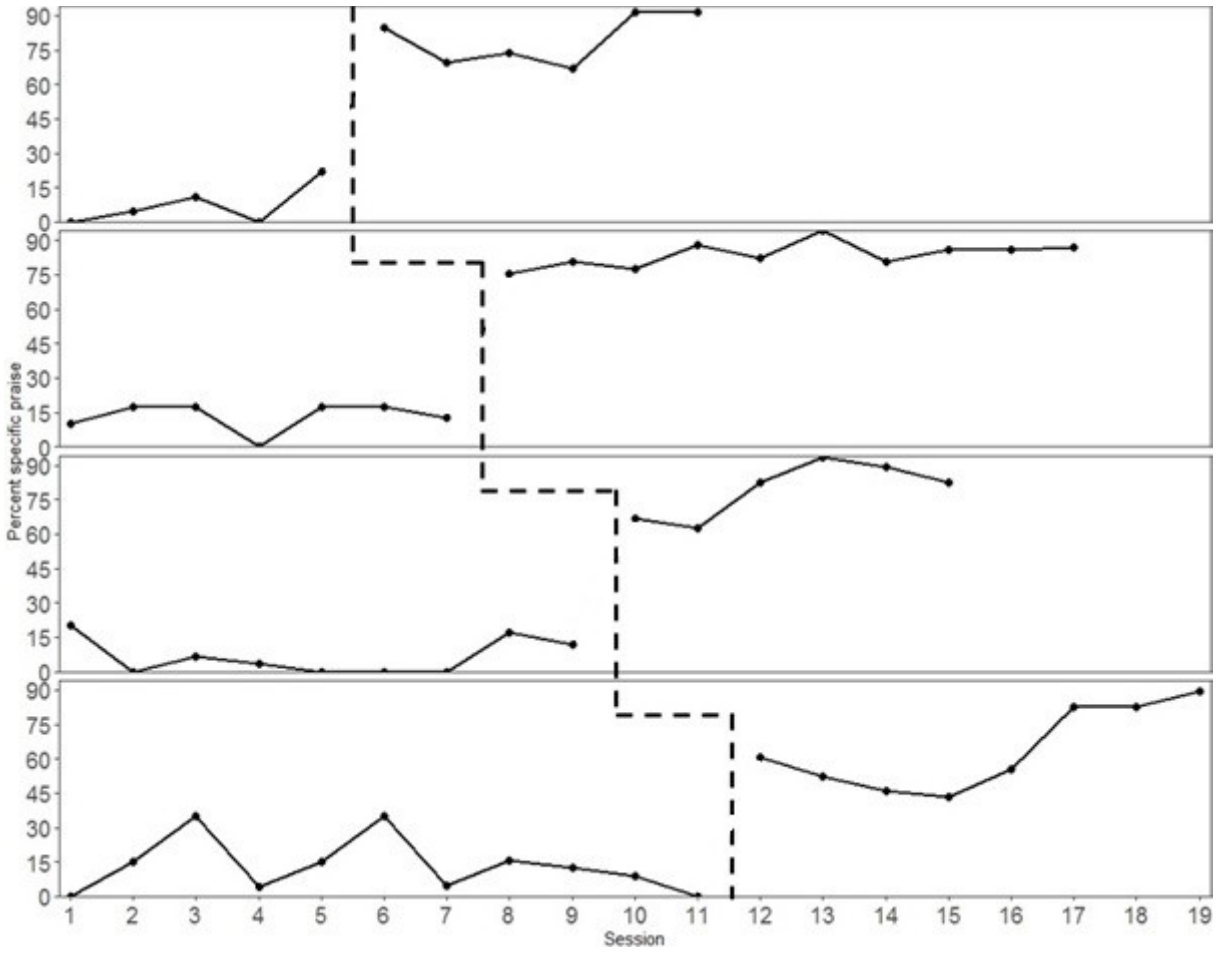

Figure 3. Example of a multiple baseline across participants design. Data from Scheeler, Morano, and Lee (2018)

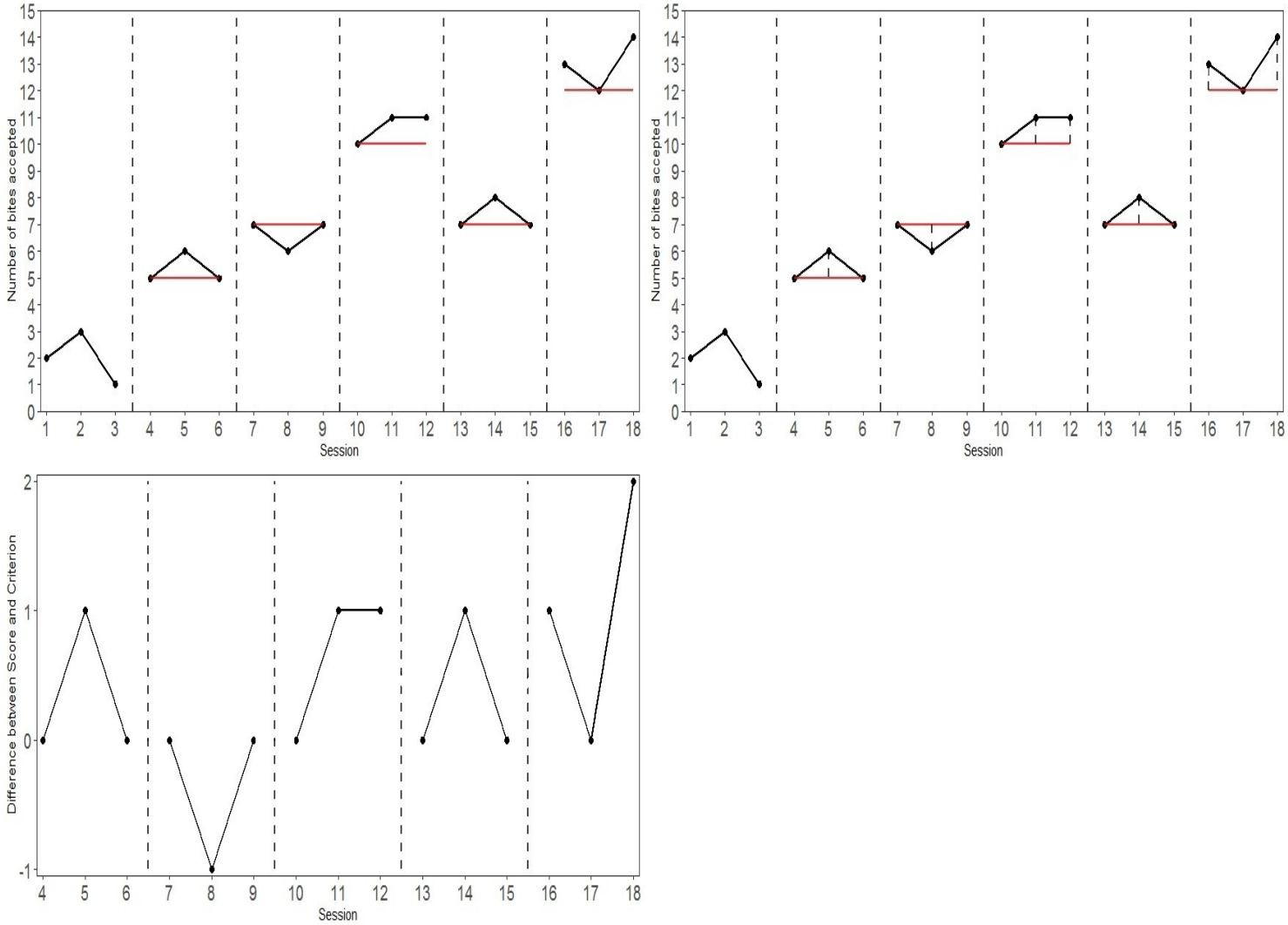

Figure 4. Example of a changing criterion design. Red horizontal lines indicate the criterion for each phase. Upper left panel: raw data; upper right panel: difference between each score and the corresponding (vertical dashed lines); lower left panel: converted difference scores from upper right panel. Data from Voulgarakis and Forte (2015).