Assessing Consistency in Single-Case Data Features

Using Modified Brinley Plots

Rumen Manolov[1] and René Tanious[2]

[1]Department of Social Psychology and Quantitative Psychology,

Faculty of Psychology, University of Barcelona

[2]Faculty of Psychology and Educational Sciences, Methodology of Educational Sciences Research Group,

KU Leuven – University of Leuven, Leuven, Belgium

***Corresponding Author:**

Rumen Manolov*

Department de Psicologia Social i Psicologia Quantitativa, Universitat de Barcelona

Passeig de la Vall d'Hebron 171, 08035, Barcelona, Spain

E-mail: rrumenovl3@ub.edu

René Tanious

Faculty of Psychology and Educational Sciences, KU Leuven

Tiensestraat 102, B-3000 Leuven, Belgium.

Phone: +32 16 32 82 19

E-mail: rene.tanious@kuleuven.be

Declarations of interest: none

Abstract

The current text deals with the assessment of consistency of data features from experimentally similar phases and consistency of effects in single-case experimental designs. Although consistency is frequently mentioned as a critical feature, few quantifications have been proposed so far: namely, under the acronyms CONDAP (consistency of data patterns in similar phases) and CONEFF (consistency of effects). Whereas CONDAP allows assessing the consistency of data patterns, the proposals made here focus on the consistency of data features such as level, trend, and variability, as represented by summary measures (mean, ordinary least squares slope, and standard deviation, respectively). The assessment of consistency of effect is also made in terms of these three data features, while also including the study of the consistency of an immediate effect (if expected). The summary measures are represented as points on a modified Brinley plot and their similarity is assessed via quantifications of distance. Both absolute and relative measures of consistency are proposed: the former expressed in the same measurement units as the outcome variable and the latter as a percentage. Illustrations with real data sets (multiple baseline, ABAB, and alternating treatments designs) show the wide applicability of the proposals. We developed a user-friendly website to offer both the graphical representations and the quantifications.

Keywords: Single-Case Experimental Design; Visual Displays; Quantitative Methods; Statistical Analysis; Experimental Replication; Research Quality

Assessing Consistency in Single-Case Data Features Using Modified Brinley Plots

Single-case experimental designs (SCEDs) enable the intensive study of one or few entities, measured repeatedly in different conditions actively manipulated by the researcher (Barlow et al., 2009; Kennedy, 2005). One of these conditions is usually a baseline, representing the usual state of affairs, whereas in the other condition an intervention takes place. In order to provide solid evidence about the impact of the intervention, it is necessary to replicate the basic effect (i.e., single A-B comparison favoring the intervention), within the same individual as in an ABAB design or across individuals as in a multiple-baseline design (Horner et al., 2005; Maggin et al., 2014; Sidman, 1960). In consequence, most of the SCED studies include several participants (Shadish & Sullivan, 2011) or reversals and there have already been specific analytical proposals for dealing with SCED studies with inter-subject replication (Ferron et al., 2009; Shadish et al., 2014).

When trying to replicate, it is crucial that the results obtained are consistent or reliable in order to be able to attribute the effect to the experimental manipulations (Ledford et al., 2019; Lobo et al., 2017; Maggin et al., 2018). Specifically, the consistency of effects has been deemed "critical" (Ledford et al., 2019, p. 36), as "consistency and replication are essential characteristics for a functional relation determination – large differences in level are not" (Lane, Ledford, & Gast, 2017, p. 7102300010p6). Moreover, according to Kazdin (2019), consistency is at the heart of what is being visually analyzed, as "visual inspection refers to reaching a judgment about the reliability or consistency of intervention effects" (p. 12). Actually, the consistency of effects is also an assumption when using a statistical procedure such as the between-case standardized mean difference (Shadish et al., 2014). Once the importance of consistency is made explicit, it is necessary to define it conceptually and operatively. Hagopian (2020) uses the term reliability to refer to the reproducibility of an effect in the same context, whereas the term consistency is used in the What Works Clearinghouse standards (Kratochwill et al., 2010). In the standards, two kinds of consistency of replications are discussed: similarity between the data patterns belonging to experimentally similar phases or conditions[1] and similarity between the effects observed when comparing across phases or conditions. The assessment of consistency has been restricted to "an overall gestalt analysis" (Geist & Hitchcock, 2014, p. 304). On the one hand, the conceptual definitions can be tautological: "consistency of data patterns across similar patterns involves looking at data from all phases of the same condition […] and examining the consistency of the data" (Kratochwill et al., 2010, p. 19). Regarding approximations to an operative definition, Maggin et al. (2013) suggest that the ratio of effects to no-effects, in the within-study replications (or attempts to demonstrate an effect), should be at least 3:1, in order to constitute evidence for an intervention effect. It could be argued that in case such a ratio is accomplished, the within-study intervention effect would be consistent enough. The question still remains how an effect is objectively demonstrated in each of the A-B comparisons, especially considering that five data features are inspected to assess the presence of an effect: differences in level, slope, and variability, immediacy, overlap (Kratochwill et al., 2010; Ledford et al., 2019; Maggin et al., 2018).

The main aim of the current text is to propose novel quantifications of consistency in similar conditions and the consistency of effects. First, existing quantifications are reviewed. Second, an emphasis is put on the difference between these existing

[1] We use "phases" when referring to designs such as withdrawal/reversal and multiple-baseline and "conditions" (as a more general term), especially necessary when dealing with alternating treatments designs in which there are no phases.

quantifications and the current proposal. Third, given that the quantifications proposed are closely related to (and easily represented on) a specific graphical representation called “the modified Brinley plot” (Blampied, 2017), we present the main features of this plot. Later in the text, we present examples of the application of the proposals. The examples are used as an initial step toward the subsequent explanations regarding to the why the present proposals have added value to existing options for assessing consistency in SCED studies. Finally, we present the quantifications formally, before deriving implications for applied researchers.

Existing Options for Quantifying Consistency

A proper operative definition of the two types of consistency is due to Tanious et al. (2020). These authors focus on ABAB designs, making two main proposals. First, they propose a quantification for the consistency of data patterns for measurements taken in experimentally similar conditions (called CONDAP). CONDAP computes the sum of the absolute differences between data points occurring at paired moments in time (e.g., the first A_1 data point is compared to the first A_2 data point). Tanious, De, et al. (2019) performed a field test and offered tentative guidelines for interpreting whether the CONDAP values suggest very high / high / medium / low / very low consistency. The application of CONDAP has subsequently been extended to multiple-baseline and changing criterion designs (Tanious, Manolov, et al., 2019). A different way of assessing consistency has been proposed for alternation designs with randomized blocks, on the basis of the way in which variability is partitioned in the analysis of variance (Manolov et al., 2020). Additionally, Tanious et al. (2020) propose a quantification of the consistency of effects (changes in level, trend, variability, overlap, immediacy) when comparing across adjacent conditions.

This quantification is called CONEFF and entails computing five effect size measures: a standardized mean difference, difference in ordinary least squares regression slopes, a variance ratio, the Nonoverlap of all pairs (Parker & Vannest, 2009) and the immediate treatment effect (Michiels & Onghena, 2019). Afterwards, the absolute difference between these effect size measures as calculated for the A_1-B_1 and A_2-B_2 comparisons is computed. CONDAP and CONEFF are applicable regardless of the length of the phases and regardless of the measurement units of the target variable.

A New Approach to Quantifying Consistency

The assessment of consistency of data patterns from experimentally similar conditions is made at the level of measurements (as in CONDAP), whereas the assessment of consistency of effects is made at the level of summary measures, representing level, trend, variability, immediacy, and overlap (as in CONEFF). In contrast, in the current text, we propose a different approach to assessing consistency, in which both the consistency in similar conditions and the consistency of effects is evaluated for summary measures[2]. To make a distinction with the quantifications in CONDAP, we refer to consistency in data *features* (as represented by the summary measures), rather than in data *patterns*. Therefore, the quantification of consistency in similar phases is proposed as complementary to (and not a substitute for) CONDAP. As an illustration of this distinction we refer to Plavnick and Ferreri (2013) who discuss the importance of finding consistent levels and trends when introducing and removing an intervention: the quantifications proposed here can be useful

[2] Overlap is not included in the current proposal because the modified Brinley plot and the consistency quantifications presented here focus on distances, whereas overlap is an ordinal comparison (data points are either overlapping or not overlapping). In addition, overlap is per definition a data aspect that is assessed between experimentally different phases. As such, overlap cannot be considered for consistency of experimentally similar conditions

for that purpose. At the same time, these authors mention that the summary measures can be similar but the underlying data patterns may be different: for assessing consistency in this way, CONDAP will be useful.

The Modified Brinley Plot

A modified Brinley plot offers the possibility to summarize, in a single graph, the results for several A-B comparisons, plotting the mean of the A phase against the mean of the corresponding adjacent B phase for each comparison. As per Blampied (2017), the word "modified" is used to indicate that means of individuals rather than groups are represented. Such plots have been suggested also for medical N-of-1 trials (Mengersen et al., 2015). Moreover, the Brinley plot is very similar to the L'Abbé plot (L'Abbé et al., 1987) for representing the results of meta-analysis in which two groups (control and treatment) are compared. The L'Abbé plot has been suggested as an exploratory tool (Sharp et al., 1996), for assessing heterogeneity (Anzures-Cabrera & Higgins, 2010; Song, 1999), similar to the use of the modified Brinley plot suggested here.

We will illustrate the modified Brinley plot with the data on context-sensitive behavioral supports for young children with traumatic brain injury obtained by Feeney and Ylvisaker (2003, 2006, 2008) in three separate studies. The aim of the studies was to reduce the frequency of challenging behaviors. Each of these studies includes an ABAB design with two participants. The time series line plots can be found in Appendix A, available at https://osf.io/uyj7w/. The modified Brinley plot includes the baseline phase means on the abscissa (X-axis) and the intervention phase means on the ordinate (Y-axis). In the plot represented in Figure 1, there are 12 points, given that there is a total of six participants and two points (i.e., two A-B comparisons) for each participant studied with an ABAB design.

Just like Tanious et al. (2020), for an ABAB design, we propose to focus only on the two comparisons that follow the AB order. Tanious et al. (2020) argue that the B_1-A_2 comparison is conceptually different, as the second baseline is preceded by an intervention, whereas the first baseline is not. Similarly, Scruggs and Mastropieri (1998) advocate for performing only the comparisons that follow the AB-sequence. It is also one of the modeling options when conducting a multilevel analysis for replicated ABAB designs (Shadish et al., 2013). Moreover, it is possible that, in case the intervention is effective, the measurements in the second baseline do not reverse to the initial baseline level or at least not immediately. Thus, we concur with Parker and Vannest (2012, p. 260), who state that "at present it seems best to omit the reversal contrast (B1 vs. A2) from effect size calculation."

INSERT FIGURE 1 ABOUT HERE

The continuous diagonal line marks the lack of change between conditions. Given that all 12 points are below this line, this entails that, for all 12 comparisons, the mean of the intervention phases is lower than the mean of the preceding baseline phase. This can be interpreted as the complete superiority of one condition over the other, in terms of mean levels. The assessment of superiority is ordinal, in the same way as it is when assessing overlap: for each comparison, it is evaluated whether the value of one condition is greater than the other or not, but not how much greater. Beyond this ordinal superiority, the farther away the points from the diagonal line, the larger the magnitude of the difference between the mean levels of the conditions.

Additional information can be represented in the modified Brinley plot. Suppose that a desired decrease is of five challenging behaviors on average. This is represented by the grey dashed diagonal line: almost all 12 points are below it, indicating that the desired reduction is achieved. Furthermore, suppose that a cut-off point can be established and that for the current example this cut-off point is 4 and is represented as a horizontal dotted line: 10 of the 12 intervention phase means are below this cut-off point. These two additional lines illustrate the fact that the modified Brinley plot can be used to represent the proportion of participants for whom a certain socially / practically desired outcome is achieved, as a recommended way of summarizing the information obtained across cases (Hagopian, 2020). This is also well-aligned with some statistical approaches for contrasting hypothesis for all participants, rather than on average (Klaassen, 2020).

Note that each individual is represented by a different color, which allows preserving information about the nested structure of the data (i.e., which A-B comparisons belong to the same individual). This is relevant, because it allows assessing visually both within-subject consistency and between-subjects consistency. For the data represented in Figure 1, there are two individuals (marked in violet and in orange) for whom the points are very close to each other, suggesting consistency in level for similar phases and also consistency of the intervention effect for each of these two participants.

Understanding Consistency in Data Features via Examples with Fictitious Data

The following three examples illustrate the concepts of consistency in data features in similar conditions and consistency of effects. All three examples represent fictitious multiple-baseline data: there is a replication of the comparison between an A phase and a B phase across three cases. For simplicity, all phases include three data points, although we are not recommending the use of such short phases and a real multiple-baseline design requires a staggered introduction of the intervention across cases. The supposed aim of the intervention is the increase of the target behavior. The focal data feature is level. Thus, when consistency in similar conditions is assessed, the means across replications of the same phase (A or B) are compared. Additionally, when consistency of effect is assessed, mean differences (obtained for each replication) are compared. The data are available at https://osf.io/uyj7w/.

Example 1: Consistency in Condition A

The time series data for Example 1 (https://osf.io/umytg/) are represented in Figure 2. It can be seen that the mean in phase A is 6 for all three tiers, although for Case 1 there is an increasing trend, for Case 2 there is an alternating pattern and for Case 3 the data are perfectly stable. The consistency in level for phase A can be clearly seen from Figure 3 (upper left panel), representing the modified Brinley plot focusing on condition A. All three baseline means are on the vertical line representing the overall mean for condition A, which is equal to 6. Given that there is no horizontal distance between the dots, the consistency in level for phase A is perfect, leading to quantifications (MAE and MAPE) equal to 0. It should be noted that the modified Brinley plot and the quantifications focus on the mean and not on the individual values or data patterns (as CONDAP does). In that sense, there is perfect consistency in level, but not perfect consistency in the data patterns, as illustrated by Figure 2.

In contrast, the consistency in condition B is not perfect, as the all phase B means (9 for Case 1, 5 for Case 2, and 8 for Case 3) are not equal to the overall phase B mean (7.33). In the modified Brinley plot (upper right panel of Figure 3), there is a vertical distance between the dots, as not all of them are on the horizontal line representing the overall phase B mean. Given that there is perfect consistency for condition A and not perfect consistency for condition B, there is also not perfect consistency of effect. The modified Brinley plot represented on the lower panel of Figure 3 shows that not all dots are on the dashed diagonal line representing the overall mean difference. Specifically, the overall mean difference is 1.33 (an increase in condition B), whereas the mean differences for each case are 3, −1, and 2. Thus, for Case 2, there is a decrease and not an increase in the target behavior.

In terms of the quantifications obtained it should be noted, for condition B, that the MAE (distance between the phase B means for each case and the overall phase B mean) is 1.56. This represents 21.21% of the overall mean for condition B (7.33). When assessing the consistency of effects, the MAE (distance between the mean differences for each case and the overall mean difference) is also 1.56. However, this value represents 116.67% of the overall mean difference (1.33), indicate that the effect is not consistent.

INSERT FIGURES 2 AND 3 ABOUT HERE

Example 2: Consistency in Condition B

The time series data for Example 2 (https://osf.io/n6ezm/) are represented on Figure 4. It can be seen that the mean in phase B is 9 for all three tiers, although for Case 1 there is an increasing trend, for Case 2 the data are perfectly stable, and for Case 3 there is an alternating pattern. The consistency in level for phase B can be clearly seen from Figure 5 (upper right panel), representing the modified Brinley plot focusing on condition B. All three intervention phase means are on the horizontal line representing the overall mean for condition B, which is equal to 9. Given that there is no vertical distance between the dots, the consistency in level for phase B is perfect, leading to quantifications (MAE and MAPE) equal to 0.

In contrast, the consistency in condition A is not perfect, as the all phase A means (6 for Case 1, 2 for Case 2, and 5 for Case 3) are not equal to the overall phase A mean (4.33). In the modified Brinley plot (upper left panel of Figure 5), there is a horizontal distance between the dots, as not all of them are on the vertical line representing the overall phase A mean. Given that there is perfect consistency for condition B and not perfect consistency for condition A, there is also not perfect consistency of effect. The modified Brinley plot represented on the lower panel of Figure 5 shows that not all dots are on the dashed diagonal line representing the overall mean difference. Specifically, the overall mean difference is 4.67 (an increase in condition B), whereas the mean differences for each case are 3, 7, and 4. Thus, for Case 2, the increase is much larger than for the remaining two cases.

In terms of the quantifications obtained it should be noted, for condition A, that the MAE (distance between the phase A means for each case and the overall phase A mean) is 1.56. This represents 35.90% of the overall mean for condition A (4.33). When assessing the consistency of effects, the MAE (distance between the mean differences for each case and the overall mean difference) is also 1.56. However, this value represents 33.33% of the overall mean difference (4.67), indicate that the effect is more consistent than in Example 1.

INSERT FIGURES 4 AND 5 ABOUT HERE

Example 3: Consistency of Effect

The time series data for Example 3 (https://osf.io/qf2cb/) are represented on Figure 6. For Case 1, there appears to be a general increasing trend throughout both conditions. For Case 2, the data are stable in both phases and the effect is clear. For Case 3, there is a decreasing trend in phase A, followed by an increasing trend in phase B with another clear effect, although in this case it combines change in level and in trend.

As shown in the upper panels of Figure 7, in this example, the mean level is not perfectly consistent either for phase A (with MAPE equal to 35.90% as a result of an overall phase A mean of 4.33 and phase A means of 6, 2, and 5 for Cases, 1, 2, and 3, respectively) or for phase B (with MAPE equal to 21.21% as a result of an overall phase A mean of 7.33 and phase A means of 9, 5, and 8 for Cases, 1, 2, and 3, respectively). In contrast, the effect, understood as difference in means, is perfectly consistency for all three cases: it is equal to 3 for all of them. In the modified Brinley plot from the lower panel of Figure 7 shows that all dots are on the dashed diagonal line. Thus, the MAE and MAPE for consistency of effect are both equal to 0.

The current example shows that the effect, measured as difference in level, can be consistent even when the level is not consistent in either of the conditions. Furthermore, Example 3 shows that the assessment of consistency cannot be performed solely on the basis of a single data feature (e.g., level), represented on the modified Brinley plot and related quantifications. It is necessary to explore the time series plot and also to take trend and variability into account when deciding whether the effect is consistent across cases or not.

INSERT FIGURES 6 AND 7 ABOUT HERE

Real-Data Examples of Assessing Consistency

In the current section we present several examples of previously published data in order to illustrate different degrees of consistency of data features in similar conditions and different degrees of consistency of effects. The data can be accessed at https://osf.io/uyj7w/, whereas the results are obtained using the websites presented with the first example.

An Example of Consistent Effects

Figure 8 represents the data gathered by Allen et al. (2015) on three behaviors (requesting help, checking out, and ordering food) exhibited by a participant called Holly, diagnosed with autism spectrum disorder and intellectual developmental disorder. Visually, there appears to be a lack of clear trend in any phase and relatively consistent levels across similar conditions and consistent change in level. The descriptive values (obtained from the numerical output of https://manolov.shinyapps.io/Brinley) are available in Table 1.

INSERT FIGURE 8 ABOUT HERE

INSERT TABLE 1 ABOUT HERE

However, it is not easy to assess consistency by only looking at the time series plot and the descriptive summaries provided. One option for summarizing the results of multiple-baseline data is via the between-case standardized mean difference (Shadish et al., 2014). Using https://jepusto.shinyapps.io/scdhlm/, the effect obtained is 4.38, 95% confidence interval [−3.30, 5.46] and an intraclass correlation (proportion of variability between cases out of the total variability) of 0, indicating that all of the variability is within cases rather than across cases. Another option for summarizing the results of multiple-baseline data is via a multilevel model (Ferron et al., 2009). Allowing for different trends in the baseline and intervention phases and modelling all effects as random, it was obtained using https://manolov.shinyapps.io/SeveralAB, based on the *nlme* package for R (Pinheiro et al., 2020).The variability in the initial baseline level and trend is rather low, whereas the immediate change in level is somewhat more variable across tiers (see Table 2 for the specific numerical results).

INSERT TABLE 2 ABOUT HERE

Quantifying the consistency of data patterns across similar phases, CONDAP (Tanious et al., 2020; computed via https://manolov.shinyapps.io/CONDAP) for phase A is 1.10 and for phase B is 1.23 – both would be labelled medium consistency according to Tanious, De, et al. (2019). Quantifying the consistency of effect as change in level via CONEFF, the differences between the phase A and phase B means are 51.52 (A_1-B_1), 35.63 (A_2-B_2), and 36.90 (A_3-B_3) (obtained from the numerical output of https://manolov.shinyapps.io/Brinley), leading to the following absolute differences in effect: |51.52 – 35.63| =15.89, |51.52 – 36.9| =14.62, and |35.63 – 36.9| = 1.27 for an average absolute difference between effects of 10.59, representing the average increase in the percentage of appropriate behavior.

To further understand the degree to which the data features and the effect are consistent, we can use the quantifications related to the modified Brinely plot. On this plot, each point represents an A-B comparison, with the phase A summary value (e.g., the mean) defining the location on the abscissa (X-axis) and the phase B summary value defining the location on the ordinate (Y-axis). It is possible to add a (larger) single point, representing the overall summary across A-B comparisons: its coordinates are defined by the overall mean of the condition A means (on the abscissa) and the overall mean of the condition B means (on the ordinate). Visual aids are added to the same configuration of dots to represent different aspects of consistency. The left panel of Figure 9 is used for assessing consistency in condition A, by drawing a vertical line denoting the overall average for condition A and marking the horizontal distance between the dots and this line. The right panel of Figure 9 is used for assessing consistency in condition B, by drawing a horizontal line denoting the overall average for condition B and marking the vertical distance between the dots and this line. Figure 10 is used for assessing the consistency of effects, by drawing a diagonal line denoting the overall effect across all A-B comparisons. This overall effect is just the mean of all differences (e.g., in level) between adjacent conditions. For a multiple-baseline design across three behaviors, the overall effect is the average of the three mean differences. To assess consistency of effects, we focus on the vertical distance between the dots for each A-B comparison and the dashed diagonal line representing the overall effect.

Figure 9 illustrates the consistency of mean level for both phase A and phase B – the smaller colored dots are very close both horizontally (phase A) and vertically (phase B) to the larger black dot representing the overall mean for each phase. For phase A, the mean absolute difference between means across tiers is only 4.44, which is 13.22% of the overall phase A mean (33.61). For phase B, the mean absolute difference between means across tiers is only 2.34, which is 3.12% of the overall phase B mean (74.96). Figure 10 illustrate the consistency of the mean difference between conditions: the mean absolute difference between effects (mean differences) is 6.78, which is 16.40% of the overall mean difference (41.35). Two of the dots are very close to the dashed diagonal line representing the overall mean difference, whereas one dot (corresponding to Tier 1) is a little bit farther away.

INSERT FIGURES 9 AND 10 ABOUT HERE

An Example of Inconsistent Effects

Figure 11 represents the data gathered by Laski et al. (1988) on verbalization in a free play context by echolalic children. Visually, the phase A levels do not appear to be consistent, whereas the phase B levels are somewhat more consistent. Consequently, the differences in level across phases are not consistent. These visual impressions are backed by the descriptive values presented in Table 1 and by the multilevel analysis results included in Table 2. Specifically, according to the multilevel results, the variability in the immediate change in level and in the change in slope of the trend lines is substantial. If only change in level is modeled, disregarding trend, the between-case standardized mean difference obtained is 1.76 (with a 95% confidence interval ranging from 1.05 to 2.47) and an intraclass correlation of 0.26 suggesting more variability across cases (as a proportion of the total variability) than for the previous data set.

INSERT FIGURE 11 ABOUT HERE

In terms of the quantifications of consistency, CONDAP yields 1.41 for phase A and 1.15 for phase B, with both values labelled as medium consistency. CONEFF applied to the change in level, calculated in the same way as for the previous example, leads to the following differences between phases: 42.54, 24.95, 4.97, and 22.00, resulting in the following absolute differences in effect: 17.59, 37.57, 20.54, 19.98, 2.95, and 17.03 for an average absolute difference between effects of 19.28 representing the percentage of target behavior. This average absolute difference between effects (i.e., lack of consistency of effects) is twice as large as in the previous example.

Providing further assessment of consistency, Figure 12 illustrates the greater consistency of mean level for phase B – the smaller colored dots are closer to the line vertically than horizontally. For phase A, the mean absolute difference between means across tiers is 7.95 (almost twice as large as the one computed for the Allen et al., 2015, data), which is 14.37% of the overall phase A mean (55.29). For phase B, the mean absolute difference between means across tiers is only 3.84, which is 4.87% of the overall phase B mean (78.91). Figure 13 illustrates the consistency of the mean difference between conditions: the mean absolute difference between effects (mean differences) is 10.13, which is 42.89% of the overall mean difference (23.62). Two of the dots are very close to the dashed diagonal line representing the overall mean difference, whereas the two remaining dots (corresponding to Tiers 1 and 3) are farther away.

INSERT FIGURES 12 AND 13 ABOUT HERE

An Example of Inconsistent Mean Level for Condition A

Figure 14 represents the data gathered by Laski et al. (1988) on verbalization in a free play context by nonverbal children (which are different from the echolalic children whose data was represented on Figure 11). Visually, the phase A levels do not appear to be consistent (given that for two tiers the level is exactly 0, whereas for two other tiers the mean level is greater than 30). The phase B levels appear to be somewhat more consistent (ranging between 47 and 60), although for Tier 2 the level is higher (78.62, as per Table 1). The differences in level seem to be relatively consistent across tiers, but a formal assessment would be necessary. According to the multilevel results presented in Table 2, there is substantial variability in the initial baseline level and also to a lesser extent in the baseline trend. Regarding the effect of the intervention, there is small variability in the immediate effect and large variability in the change in slope. If only change in level is modeled, disregarding trend, the between-case standardized mean difference obtained is 2.13, 95% confidence interval [1.04, 3.22] and an intraclass correlation of 0.50 suggesting more variability across cases (as a proportion of the total variability) than for the previous two data sets.

INSERT FIGURE 14 ABOUT HERE

In terms of the quantifications of consistency, CONDAP yields 2.46 for phase A (very low consistency) and 1.46 for phase B (medium consistency). CONEFF applied to the change in level leads to the following differences between phases: 29.04, 44.01, 49.41, 47.76, and 39.31, leading to the following absolute differences in effect: 14.97, 20.37, 18.72, 10.27, 5.40, 3.75, 4.70, 1.65, 10.10, and 8.45 for an average absolute difference between effects of 9.84 representing the percentage of target behavior. In general, comparing CONEFF across studies, when calculated based on the raw mean, is reasonable only when the measurement units are the same. Given that both the Allen et al. (2015) data (Figure 8) and the two data sets from Laski et al. (1988) represented on Figures 11 and 14, use percentages, such a comparison is justified. For the Laski et al. data on nonverbal children, the effect is slightly more consistent than for the Allen et al. data.

Providing further assessment of consistency, Figure 15 illustrates (via larger horizontal and vertical distances between the dots) the smaller consistency of mean levels for similar phases, as compared to the previous data sets. For phase A, the mean absolute difference between means across tiers is 13.69 (more than three times as large as the one computed for the Allen et al., 2015, data), which is 80.00% of the overall phase A mean (17.11). For phase B, the mean absolute difference between means across tiers is only 8.46

(twice as large as the one computed for the Allen et al., 2015, data), which is 14.14% of the overall phase B mean (59.01). Figure 16 illustrates that the consistency of the mean difference between conditions is similar to the one for the Allen et al. (2015) and larger than the one observed for the echolalic children. The mean absolute difference between effects (mean differences) is 6.19, which is 14.76% of the overall mean difference (41.90). Four of the dots are relatively close to the dashed diagonal line representing the overall mean difference, whereas one dot (corresponding to Tier 1) is farther away.

INSERT FIGURES 15 AND 16 ABOUT HERE

The Need for the Modified Brinley Plot and Related Quantifications for Assessing Consistency

The Graphical Representations Allow for an Efficient Representation of the Data Features and the Quantifications

It is useful to summarize the information obtained when several participants take part in the same study, making possible the comparison across individuals. Hagopian (2020) advocates for tabular representations of the main features of the design and the outcome. Another option is the visual summary that a modified Brinley plot can provide. The summary measures (means, slopes of trend lines, standard deviations) are likely to be more reliable when more measurements per phase are available, which is also well-aligned with the requirements of having multiple measurements per phase (e.g., at least five, Ganz & Ayres, 2018).

It can be argued that it is cognitively challenging to interpret all visual information represented in the traditional time series line graphs, especially if consistency is to be assessed. In fact, to the best of our knowledge, there is no specific proposal for the graphical depiction of consistency applicable to all kinds of SCEDs[3]. If time series graphs are to be used for visually assessing consistency, it is first necessary to distinguish consistency in similar phases or conditions from consistency of effects. Second, the visual assessment of consistency requires inspecting and keeping in the working memory several graphical representations at a time. Moreover, this visual assessment is also not easily translated into a clearly communicable outcome. Quantifications (such as CONDAP, CONEFF, and the proposals made here) can make this task easier, but can also be easier interpreted and understood in combination with a specific graphical representation. This is achieved via the modified Brinley plot: several A-B comparisons for several participants can be represented on the same plot, with vertical and horizontal distances representing the degree of lack of consistency. These distances are the basis of the quantifications proposed and thus the quantifications and the graphical representations are intimately related.

In terms of efficiency in summarizing information about consistency in similar conditions, CONDAP yields a single number after comparing data patterns, whereas the proposals made (MAESIM and MAPESIM) here provide one quantification for each data feature: level, trend, and variability. Regarding the consistency of effects, the quantifications underlying CONEFF have not yet been made completely automatic, whereas the proposals made here (MAEDIFF and MAPEDIFF) are implemented in software providing automatically the quantification and a graphical representation of how close each effect is to the overall effect.

[3] See Manolov et al. (2020) for a proposal applicable to alternating treatment designs with block randomization.

The Proposals Are Intended to Complement Not Substitute

Visual Representations

As illustrated by the previous examples, the modified Brinley plot is to be used together with the common time series line plot. Inspecting the sequence of all data points is necessary for a better understanding of the process studied (Fahmie & Hanley, 2008). Moreover, it is important not to discard detailed information about each measurement obtained for each participant, as it is expected to be provided in a tabular or a graphical form (Tate et al., 2016). Apart from containing raw data, the time series line graph is especially necessary for visually assessing the degree to which the mean or the trend line is a reasonable representation of the data within a phase, as there may be excessive variability or outlying data points, both of which can affect the meaningfulness of summary measures such as the mean and the slope of the trend line.

In relation to multilevel models, it is possible to represent graphically the level-2 model fitted to the data and can be represented alongside level-1 trend lines to visually inspect the degree to which the averages represent the individual data (Declercq et al., 2020). Nonetheless, beyond a visual impression, this graphical representation is more difficult to link directly to a quantification of consistency across participants. The modified Brinely plot and the related quantifications provide this direct link, given that what is quantified are the distances between the dots (e.g., means) in the plot.

Data Analysis

The proposals focus on the assessment of consistency and, thus, they are not intended to provide effect size estimates or to substitute data analytical techniques with a more general purpose. The illustrations provided show how the quantifications of consistency can be used jointly with other techniques for SCED data analysis (e.g., between-case standardized mean difference, multilevel models[4], or any quantification of effect size, among the multiple options available (e.g., Busse et al., 2015; Manolov & Moeyaert, 2017) and for the assessment of consistency (i.e., CONDAP and CONEFF). The convergence of conclusions of different techniques is useful for increasing the confidence in these conclusions (Lobo et al., 2017).

Quantifications of Variability across Cases in the Context of Data Analysis Techniques

When discussing the need for a novel proposals for assessing consistency, we could look for quantifications of the variability across cases that are part of a SCED data analytical procedures applicable to multiple-baseline designs and to replicated withdrawal/reversal designs. Such a quantification is the intraclass correlation, which is provided in the output of the between-case standardized mean difference. Specifically, it is used in the small-sample correction and for estimating the variance of the effect size (Shadish et al., 2014). However, the intraclass correlation is not a pure quantification of (lack of) consistency in similar phases, or of consistency of effects. This is so because the intraclass correlation can be equal to zero even when the mean differences are not exactly the same for all participants and/or the means in each phase are not exactly the same for all participants (as in the example provided with the data by Allen et al., 2015). Moreover, the intraclass correlation increases in two different scenarios: (a) when there is a difference across cases

[4] The between-case standardized mean difference and multilevel models can be understood as mainly focussing on level as a data feature (Chen et al., 2015), whereas the current proposal focuses on consistency, using the taxonomy of data features provided by Kratochwill et al. (2010).

in terms of the level change between adjacent phases and (b) when the mean level within each phase is different across cases, even when the level change is the same.

In relation to multilevel models, this analytical option provides both a quantification of the average effect across cases and a quantification of the variability across cases. This quantification of variability can be expressed as a standard deviation in the initial baseline level, in baseline trend and in the immediate difference in level or the difference in slope. It should be noted that, unless the baseline trend is zero, the initial baseline level (i.e., the value at the beginning at the baseline) is not the same as the mean baseline level. The latter is the data aspect assessed in the current proposal; it can be obtained from a multilevel model not including trend. Similarly, the immediate effect of the intervention is not the same as the difference between the means of the two phases, which is what is assessed in the current proposal and also requires a multilevel model not including trend. Moreover, there is not a direct quantification of the consistency of the data features in phase B. An additional issue related to the random effects or variance estimates provided by multilevel models is that they are biased (Baek et al., 2020; Ferron et al., 2009; Moeyaert et al., 2017, negatively bias in the case of using maximum likelihood estimation) and thus are likely to be inappropriate as quantifications of the degree of inconsistency across cases.

Another quantification from multilevel analysis that could be considered as potentially useful for assessing consistency is the intraclass coefficient. In intercept-only models (also referred to as null, unconditional or empty model, Ferron et al., 2008, i.e., not modelling trend or changes in relation to the different phases), the intraclass correlation can be interpreted as the proportion of variance explained by the nesting structure, with measurements nested into participants, or the correlation between measurements within participants (Hox, 2010). The intraclass correlation is not commonly computed, or used for assessing consistency, once a full model is built (i.e., introducing a dummy variable representing the phase and a time variable representing session number) for quantifying differences across phases, because its calculation and interpretation is no longer straightforward when predictors are include in the model (Goldstein et al., 2002). Instead, the intraclass correlation it commonly apply to a null model (Merlo et al., 2006) and used for deciding whether a multilevel model is necessary: the larger the intraclass correlation, the more evidence that there are differences between cases, and thus the nesting structure of the data needs to be taken into account (Gage & Lewis, 2014; Peugh, 2010). It can also be useful for a priori power analysis (Shadish et al., 2013).

In summary, the proposals made here may be used together with the quantifications of variability across cases that are part of data analytical procedures, because they do not provide the same information.

Specific Quantifications of Consistency

In relation to the previous proposals for quantifying consistency, CONDAP (Tanious et al., 2020) focuses on comparing the measurements themselves, not the means or other summary measures (slopes or standard deviations). Thus, CONDAP is capable of assessing the consistency of data patterns within similar phases, but it does not inform whether the means are similar, whether the immediate effect is similar, whether the slopes are similar, or whether the variability is similar. In contrast, the modified Brinley plot and the MAPESIM focus on such summary measures representing these four data features. Moreover, by performing the calculations on the summary measures, MAPESIM performs the assessment of consistency in similar phases at the same level as the assessment of consistency of effects (which does not focus on data patterns).

In relation to visual inspection, the modified Brinley plot entails a more direct and straightforward visual representation of the consistency in the mean level of phases representing the same condition. Moreover, the calculations underlying MAPESIM are potentially more straightforward than the CONDAP quantifications. However, the easier visual inspection comes with the price of losing information about the data pattern and the need to rely on summary measures which may be distorted by outliers.

Both CONEFF (Tanious et al., 2020) and MAPEDIFF can be used to compare changes in level, in slope, and in variability. One difference is that for the change of variability as an effect, CONEFF uses the variance ratio, whereas on a modified Brinley plot the difference between standard deviations is represented and quantified. As a limitation, the modified Brinley plot and MAPEDIFF are not applicable to assess consistency in nonoverlap, as it is a single dimension (i.e., there is not one measure for phase/condition A and another measure for phase/condition B). An advantage of MAPEDIFF over CONEFF is that the latter is easily represented graphically on the modified Brinley plot.

In summary, the current proposals can be used together with CONDAP and CONEFF, given that the information provided is not identical. In that sense, if both approaches to the assessment of consistency agree, the researchers would be more confident that there results can be (or cannot be) considered consistent.

The Proposals are Applicable to Different Single-Case Experimental Designs

The initial example provided here is from multiple-baseline design studies. However, the graphical representations and quantifications suggested here are also applicable to withdrawal/reversal designs, with or without replication across participants. For an ABAB design for a single participant, there would be two dots represented on the modified Brinley plot: one for each A-B comparison. For an ABABAB across two participants, there wuld be three dots for each of the participants, for a total of six dots in the modified Brinley plot. See Appendix A (https://osf.io/uyj7w/) for an example of three studies (an original and two replications) using withdrawal/reversal designs (Feeney & Ylvisaker, 2003, 2006, 2008).

For an alternating treatments design (ATD), it is also possible to use a modified Brinley plot, although the application is somewhat different for the different kinds of ATDs (see Onghena & Edgington, 2005, for a discussion). The most direct application is with randomized block designs, also referred to as ATDs with block randomization (Ledford, 2018). This kind of ATD corresponds to the situation in which there is “random alternation with no condition repeating until all have been conducted” Wolery et al., 2018, p. 304). In an ATD with two conditions using block randomization, within each block there is one measurement in condition A and one measurement in condition B. Comparisons of treatment efficacy should thus be performed within blocks. This randomly determined sequence is equivalent to the N-of-1 trials used in the health sciences (Nikles & Mitchell, 2015), where the several random-order AB blocks are called multiple crossovers. It is possible (and in the health sciences common) to replicate the series across several participants, each with its own randomly determined sequence. In the simplest case, in the modified Brinley plot for a single participant, there would be one point for each block in the sequence. Thus, there are no means involved. The only information lost, in comparison to a time series line graph is about the temporal sequence of the measurements. See Appendix B (https://osf.io/uyj7w/) for an example with the data gathered by Lloyd et al. (2018). Another common version is the ATD with restricted randomization or restricted ATDs (Onghena & Edgington, 1994), with the usual requirement of no more than two consecutive administrations of the same condition (Heyvaert & Onghena, 2014; Wolery et al., 2018). With a restricted ATD, it is possible to obtain more sequences than with an ATD with randomized blocks. For instance, a sequence such as AABBAABBAB can be obtained. For such a sequence, the researcher would have to choose which comparisons to represent on the modified Brinley plot: three comparisons (segmenting the sequence as AABB-AABB-AB) or four comparisons (AAB-BA-ABB-AB). Therefore, for some sequences there would be more than one modified Brinley representation of the same data. The main issue with the use of the modified Brinley plot (and the quantifications of consistency presented later) is that the unit of analysis presented on the plot is not clear. In both reversal and multiple-baseline designs each point represents a summary measure of two adjacent phases, whereas in an ATD with block randomization each point represents the two measurements (representing two different conditions) in the same block. However, in a restricted ATD the exact meaning of a segment is not clear. For that reason, we consider that the modified Brinley plot (and the quantifications of consistency proposed in the current text) are not easily applicable.

Regarding changing criterion designs, the assessment of effects is usually performed in terms of whether the actual measurement correspond closely to the predefined criterion levels (Ledford & Gast, 2014; McDougall et al., 2006). Therefore, we consider that the modified Brinley plot would not be useful.

The Proposals Are Applicable beyond the Focus on Overall Mean Levels

The points on the Brinley plot need not represent the means of all measurements in the same phase in ABAB and multiple-baseline designs. In terms of level, the focus could be put on the last three baseline data points and the first three intervention phase data points: in this case, the modified Brinley plot would represent immediate effects, as defined in Kratochwill et al. (2010) and Michiels and Onghena (2019). Whether immediacy is expected (and thus relevant) or not depends on the type of intervention and target behavior (Kratochwill et al., 2014). It is also possible to use only the last three data points in each phase, as in the percentage change index (Hershberger et al., 1999; Olive & Smith, 2005). Such use of later measurements aligns well with the idea to switch conditions (Kazdin, 1978) or to estimate level when stability is reached (Ferron et al., 2020) or when looking for delayed effects (Levin et al., 2017).

Despite the fact that modified Brinley plots are typically based on using means (Blampied, 2017), the points can represent summary measures of trend or variability. Specifically, each point could also be defined by the slope of the baseline trend and the slope of the intervention phase trend, using one of the alternatives for fitting trend lines to single-case data (Manolov, 2018). If we aim for continuity with CONEFF (Tanious et al., 2020), the slope can be estimated via ordinary least squares estimation. Otherwise, the Theil-Sen method can be used (Vannest et al., 2012), because it is more resistant to outliers. Similarly, each point could represent a measure of variability. As in CONEFF (Tanious et al., 2020), the standard deviation can be used, although other measures are possible, such as the median of the absolute deviations from the median (MAD). Using the Laski et al. (1988) data, Appendix C presents the results of assessing visually and quantitatively consistency for trend, whereas Appendix D presents the same assessment for variability. Both appendices can be accessed from https://osf.io/uyj7w/. Furthermore, the consistency of the immediate effect can also be studied via the website developed (https://manolov.shinyapps.io/Brinley).

In summary, the modified Brinley plot and the quantifications presented can be used to assess consistency in similar phases (for ABAB designs and MBDs) in terms of overall level, trend and variability and the consistency of effect understood as difference in slope or in variability, apart from consistency in immediate or overall effect.

The Proposals are Applicable for Assessing Consistency within Cases and Consistency across Cases

The graphical representations and the quantifications can be applied for studying the degree of consistency of data features and the degree of consistency of effects for a single participants (e.g., in a withdrawal/reversal design) or for several participants simultaneously (e.g., in a multiple-baseline design). Specifically, for a study using a multiple-baseline design across four participants (e.g., Ciullo et al., 2015), there will only be an assessment of consistency across cases because each tier receives each treatment only once. For a study using a multiple-baseline design that is both across four participants and three contexts (e.g., Koegel et al., 1992), there can be an assessment of consistency (a) across contexts, for each individual separately, (b) across individuals, for each context separately, and (c) using all data simultaneously. According to the data included in the input data file, the website created can represent graphically and quantify consistency, as described previously, for each of these situations. Similarly, for an ABABAB design replicated across three cases (e.g., Angell et al., 2011), there can be a visual and quantitative assessment of consistency using all the data or focusing on each case separately. The same is the case for an alternation design, replicated across cases (e.g., Fletcher et al., 2010). If summary measures for several participants are represented on the same modified Brinley plot, they can be marked with different colors, like in the examples provided in Appendix A.

The Proposals Are Implemented in a User-Friendly Freely Available Website

The website developed implementing the proposals (https://manolov.shinyapps.io/Brinley) requires a data file to be uploaded and several specifications to be chosen, simply by clicking. These specifications include the aim of the intervention (to increase or reduce the target behavior), the type of design (using phases as in a multiple-baseline or reversal/withdrawal design or rapidly alternating conditions as in an ATD), and the data feature to represent (level summarized via the mean, variability measured via the standard deviation or trend estimating using ordinary least squares estimation).

The output of the software includes a time series graph and several modified Brinley plots (for assessing the consistency in each condition and for assessing the consistency of effect). Furthermore, descriptive statistics for each condition and the quantifications of consistency are obtained in the “Quantifications” tab.

Finally, note that with the modified Brinley plot it is possible to assess visually the clinical significance of the change, in case there is a substantive criterion available or sufficient previous research to inform the specification of minimally important effect. Specifically, it is possible to specify a critical value of improvement that is desired to be achieved in condition B: such a value would be represented as a horizontal dotted line.

Moreover, the user can represent the magnitude of meaningful change from baseline. This meaningful change can be expressed in absolute terms (e.g., a decrease of 5 points) or in relative terms (e.g., a decrease of 25% in comparison to the baseline level). The magnitude of meaningful change is represented as a dashed diagonal line, which is parallel to the main solid diagonal line indicating lack of effect for changes specified in absolute terms (e.g., a difference of 5 behaviors). In case the magnitude of meaningful change is specified in relative terms (e.g., a 25% reduction with respect to the baseline mean), the dashed diagonal line would not be perfectly parallel to the main solid diagonal line – the former would be closer to the latter for lower values of the baseline mean and farther away for larger values.

The use of the website is described in several documents, linked in the website itself: https://manolov.shinyapps.io/Brinley, in the last tab called "Tutorials".

Formal Presentation of the Quantifications of Consistency

All quantifications presented here are actually quantifications of lack of consistency, because larger values represent greater differences from the reference value (e.g., the overall mean). In that sense, a value of 0, for the absolute measures, or 0%, for the relative measures, would imply perfect consistency.

Assessment of Consistency of Data Features in Similar Phases or Conditions

Visually, focusing on the consistency of level of the A phases, the closer the points are horizontally, the greater the consistency. Regarding the consistency of level of the B phases, the closer the points are vertically, the greater the consistency. Therefore, it is necessary to separately evaluate the two dimensions of the plot: the abscissa (X-axis) for the consistency of data features for the A condition and the ordinate (Y-axis) for the consistency

For the assessment of consistency in similar conditions, a quantification can focus on the difference with respect to the overall mean for a given condition. We refer to this quantification as MAESIM: Mean Absolute Error for SIMilar phases or conditions. For condition A, for which the relevant dimension is the X-axis (or horizontal distance), the quantification would be

$$MAESIM = \frac{\sum_{j=1}^{J}\sum_{i=1}^{I}\left|\bar{x}_{A(ij)} - \bar{x}_{A(..)}\right|}{(I \times J)}, \qquad (1)$$

where J is the number of individuals, I is the number A-B comparisons per individual, $\bar{x}_{A(ij)}$is mean for condition A for the *j*-th individual and the *i*-th comparison, $\bar{x}_{A(..)}$is the mean for condition A across all individuals, and $(I \times J)$ is the total number of A-B comparisons. For condition B, the relevant dimension is the Y-axis (or vertical distance) and the formula can be expressed using the symbols $\bar{y}_{B(ij)}$ and $\bar{y}_{B(..)}$, but the calculation is otherwise exactly the same.

For instance, for an ABAB design replicated across three individuals, $J = 3$ and $I = 2$, whereas for a multiple-baseline design across four individuals $J = 4$ and $I = 1$. For an ATD with 6 blocks, replicated across two individuals, $J = 2$ and $I = 6$. This quantification is equivalent to the average absolute deviation or to the mean absolute error.

MAESIM is expressed in raw units, that is, in the same measurement units as the dependent variable. This can be useful for making the quantifications more easily interpretable within each study. Nonetheless, raw units preclude comparisons of consistency across studies using different operative definitions or measurement units for the same target behavior or outcome. To make the quantification relative, we propose expressing the difference between each value and the mean for the condition to which this value belongs as a percentage. This would entail using the Mean Absolute Percentage Error for SIMilar conditions or phases (Hyndman & Koehler, 2006) instead of the mean absolute error, leading to the MAPESIM quantification, here illustrated with the subscript A, for condition A:

$$MAPESIM = \frac{\sum_{j=1}^{J}\sum_{i=1}^{I}\left|\bar{x}_{A(ij)} - \bar{x}_{A(..)}\right| / \bar{x}_{A(..)}}{(I \times J)} \times 100 \qquad (2)$$

Assessment of Consistency of Effects

Visually, in order to state that there is a consistent improvement, rather than a consistent deterioration, the data points should be on the correct side of the diagonal line in the modified Brinley plot (i.e., above it, if an increase is required and below it, if a reduction is required). Additionally, the average change in level can be represented as a dashed diagonal line, parallel to the line of no change. Note that if all the individual points are placed on the dashed diagonal line representing the average effect, this would suggest perfect consistency of effect. Nonetheless, this would not entail perfect consistency in similar phases, as long as there is any horizontal or vertical distance between each individual point and the overall average point.

The quantification of consistency of effect can focus on the difference with respect to the overall A-B mean difference. For an ABAB design, this would entail calculating the difference between the first A-phase and the first B-phase (e.g., in terms of level), then executing the same calculation for the difference between the second A-phase and the second B-phase, and subsequently calculating the average of these two values. We refer to this quantification as MAEDIFF: Mean Absolute Error for DIFFerences between phases or conditions.

$$MAEDIFF = \frac{\sum_{j=1}^{J}\sum_{i=1}^{I}|d_{ij}-d_{..}|}{(I \times J)}, \quad (3)$$

where d_{ij} is the difference between the A mean and the B mean for the *ij*-th comparison (or the difference between the A measurement and the B measurement for the *ij*-th block in an ATD with randomized blocks) and $d_{..}$ is the overall mean difference. The quantification is equivalent to an average absolute deviation or a mean absolute error. To make MAEDIFF relative, MAPEDIFF can be computed, following the logic of MAPESIM. Specifically:

$$MAPEDIFF = \frac{\sum_{j=1}^{J}\sum_{i=1}^{I}|d_{ij}-d_{..}|/|d_{..}|}{(I \times J)} \times 100 \quad (4)$$

Discussion

The current text aimed to develop a proposal for assessing consistency in data features in similar phases and consistency of effects for SCEDs. The proposal was designed to integrate the visual and quantitative information, so that the calculations performed represent directly what is visible on the plot. The ease of understanding the quantifications and ease of obtaining them via a freely available website is likely to make the more attractive to applied researchers, who are still only rarely using sophisticated statistical techniques (Jamshidi et al., 2017; Natesan, 2019; Radley et al., 2020). The use of the new proposals was illustrated with published data sets and their applicability was demonstrated alongside other data analytical techniques, and other quantifications of consistency were discussed. It was shown that the new proposals add information beyond existing techniques and can be used together with other visual and quantitative SCED data analytical approaches.

Implications for Applied Researchers

Joint Visual and Quantitative Assessment of Consistency

The modified Brinley plot of phase means (for ABAB designs and MBDs) or of measurements within blocks (for ATDs with block randomization) can be used for visually representing the summary measures of several participants and several A-B comparisons per participant. The quantifications proposed are unambiguously represented on the modified Brinley plots as vertical or horizontal distances. In that way, it is possible to use visual analysis as a way of validating numerical results (Parker et al., 2006), even when assessing consistency. Thus, the compact visual representation comes with an easily communicable summary, such as MAESIM or MAEDIFF in absolute terms, and MAPESIM and MAPEDIFF, in relative terms. Both the quantifications and the graphical representations can be easily obtained via the website developed for that purpose: https://manolov.shinyapps.io/Brinley/. Thus, the current proposals are an adequate illustration of the frequently suggested joint use of visual and statistical analysis (Fisher et al., 2003; Harrington & Velicer, 2015; Karazsia, 2018).

Assessing Consistency Within or Across Participants

Within the same study using the same operative definitions and measurement units for the target behavior, consistency can be assessed jointly for all participants using either the absolute or the relative measures. Moreover, these measures can also be used for comparing between participants when using replicated ABAB designs or ATDs. Using the Feeney and

Ylvisaker data (2003, 2006, 2008, Appendix A) we illustrated how the consistency quantifications can be applied jointly to replication studies. Additionally, using the relative measures, the consistency of the average effect can be compared across different single-case studies assessing the same intervention.

Is Consistency Necessary or Sufficient?

It is desirable that data from similar conditions are consistent in level, trend, and variability (Kratochwill et al., 2010). In that sense, this type of consistency is necessary. It is not sufficient for demonstrating intervention effectiveness, because an assessment of whether the effect is consistent (within a participant or across participants) is also necessary. However, consistency of effects may not always be expected to take place. For instance, in an MBD across behaviors, lack of consistency in the effect of the intervention may stem from the behaviors not being sufficiently functionally similar and independent (Tate & Perdices, 2019).

Even for an MBD across participants or an ABAB design, it has to be considered whether it is reasonable to expect a consistent effect in terms of all these data features: level, trend, and variability. Following the discussion by Kavelaars (2020), the decision rule about superiority when there are multiple outcomes, it is necessary to establish the decision rule when multiple data features are assessed. A requirement could be to observe a consistent effect for (a) the most important data feature, (b) all data features, (c) any data feature, ignoring the remaining ones, or (d) any data feature, in case its effect is large enough to override any potential null or smaller negative effects for the remaining data features. We consider that in case there are solid expectations regarding the type of effect of the intervention, the focus could be put on the data feature that is considered to represent best this expected effect (just like choosing a test statistic according to the effect expected when using a randomization test; Heyvaert & Onghena, 2014). Otherwise, option (d) appears to be reasonable.

While consistency is crucial for increasing the confidence in the effectiveness of an intervention, lack of consistency can also be informative. Specifically, in case the results across participants are inconsistent, any difference in the features of the participants themselves, in the target behavior or in the intervention can be used as moderators for understanding better the situations in which the intervention can be expected to be most beneficial (Ledford et al., 2016; Romeiser-Logan et al., 2017).

It should be noted that the consistency of data in similar phases and consistency of effects, even when deemed to be present to a considerable degree, are not sufficient. Specifically referring to the consistency of effects, the effect should be large enough to be of practical importance (Horner et al., 2005; Kazdin, 1997), such as reaching the level of the target behavior exhibited by a normative sample of individuals (Ferron et al., 2020). This requirement brings us back to the possibility to represent, on the modified Brinley plot, the desired level of the target behavior once the intervention is in place.

Even if an effect is deemed consistent, thanks to CONEFF and/or the quantifications proposed here, it has to be kept in mind that these quantifications refer to the data actually obtained, usually at the within-study level. Thus, these quantifications can be useful as information for assessing the degree of internal validity in relation to the importance of observing consistent effects for inferring causality (Romeiser-Logan et al., 2017) or defining a successful replication (Leek & Jager, 2017). Still, it may not be directly obvious how many replications are necessary (e.g., Cook et al., 2015, suggest at least 75% of correct replications), whether small consistent effects are sufficient, or whether a large effect can be considered to make further replications unnecessary (Lanovaz et al., 2019). In any case, the degree of generalization of findings to other participants beyond the ones included in a given study (i.e., assessment of external validity) depends on logical arguments in relation to the similarity, across studies, between participants, target behaviors, and interventions.

Limitations and Future Research

The current text presents the rationale of the use of modified Brinley plots to assess consistency of data features in similar conditions and also consistency of effects. Quantifications in absolute and relative terms are also presented. For MBDs and ABAB designs, both the graphical representations and the quantifications are based on summary measures rather than on the individual measurements. In that sense, we consider that researchers willing to assess the consistency of data patterns (rather than means or slopes) for similar phases, should complement the current proposals with CONDAP (Tanious et al., 2020).

The quantifications proposed are purely descriptive – they refer to the data actually obtained. No sampling distribution has been derived for making possible the construction of confidence intervals or statistical testing. Nevertheless, the reference or inference to a population (via confidence intervals and *p*-values) may not be justified in absence of random sampling and, therefore, any developments in terms of confidence intervals or statistical significance need to stem from a design feature such as the presence of randomization (Tanious, De, & Onghena, 2019). If randomization is present in the design, then the quantifications presented here can be calculated for all possible randomizations, and the actually observed value can then be located in the randomization distribution to obtain a *p*-value and/or confidence intervals (Michiels, Heyvaert, Meulders, & Onghena, 2017).

We have presented the graphical representations and quantifications via illustrations with real data from different kinds of SCED. Nonetheless, the text does not present the results of a broader application to simulated or real data. Specifically, a field test would be necessary to provide interpretative benchmarks for the relative quantifications of consistency (MAPESIM and MAPEDIFF), following the example of Tanious, De, et al. (2019). Moreover, the shape of the distribution of MAPESIM can be compared to the one obtained for CONDAP. It would be especially important to perform field tests separately for the different design structures, because for ATDs with block randomization the points represent individual measurements, rather than means and, for that reason, they can be expected to be less consistent. In that sense, different interpretative benchmarks are likely to be necessary for ATDs with block randomization.

References

Allen, K. D., Vatland, C., Bowen, S. L., & Burke, R. V. (2015). An evaluation of parent-produced video self-modeling to improve independence in an adolescent with intellectual developmental disorder and an autism spectrum disorder: A controlled case study. *Behavior Modification, 39*(4), 542–556. https://doi.org/10.1177/0145445515583247

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*(4), 206–217. https://doi.org/10.1177/1088357611421169

Anzures-Cabrera, J., & Higgins, J. P. (2010). Graphical displays for meta-analysis: an overview with suggestions for practice. *Research Synthesis Methods*, *1*(1), 66-80. https://doi.org/10.1002/jrsm.6

Barlow, D., Nock, M., & Hersen, M. (2009). *Single case experimental designs: Strategies for studying behavior change* (3rd Ed.). Allyn and Bacon.

Baek, E., Beretvas, S. N., Van den Noortgate, W., & Ferron, J. M. (2020). Brief research report: Bayesian versus REML estimations with noninformative priors in multilevel single-case data. *The Journal of Experimental Education, 88*(4), 698–710. https://doi.org/10.1080/00220973.2018.1527280

Blampied, N. M. (2017). Analyzing therapeutic change using modified Brinley plots: History, construction, and interpretation. *Behavior Therapy*, *48*(1), 115–127. https://doi.org/10.1016/j.beth.2016.09.002

Busse, R. T., McGill, R. J., & Kennedy, K. S. (2015). Methods for assessing single-case school-based intervention outcomes. *Contemporary School Psychology, 19*(3), 136–144. https://doi.org/10.1007/s40688-014-0025-7

Chen, L.-T., Peng, C.-Y. J., & Chen, M.-E. (2015). Computing tools for implementing standards for single-case designs. *Behavior Modification, 39*(6), 835–869. https://doi.org/10.1177/0145445515603706

Ciullo, S., Falcomata, T. S., Pfannenstiel, K., & Billingsley, G. (2015). Improving learning with science and social studies text using computer-based concept maps for students with disabilities. *Behavior Modification, 39*(1), 117–135. https://doi.org/10.1177/0145445514552890

Cook, B. G., Buysse, V., Klingner, J., Landrum, T. J., McWilliam, R. A., Tankersley, M., & Test, D. W. (2015). CEC's standards for classifying the evidence base of practices in special education. *Remedial and Special Education, 36*(4), 220–234. https://doi.org/10.1177/0741932514557271

Declercq, L., Cools, W., Beretvas, S. N., Moeyaert, M., Ferron, J. M., & Van den Noortgate, W. (2020). MultiSCED: A tool for (meta-)analyzing single-case experimental data with multilevel modeling. *Behavior Research Methods, 52*(1), 177–192. https://doi.org/10.3758/s13428-019-01216-2

Fahmie, T. A., & Hanley, G. P. (2008). Progressing toward data intimacy: A review of within-session data analysis. *Journal of Applied Behavior Analysis, 41*(3), 319-331. https://doi.org/10.1901/jaba.2008.41-319

Ferron, J. M., Bell, B. A., Hess, M. R., Rendina-Gobioff, G., & Hibbard, S. T. (2009). Making treatment effect inferences from multiple-baseline data: The utility of multilevel modeling approaches. *Behavior Research Methods, 41*(2), 372–384. https://doi.org/10.3758/BRM.41.2.372

Ferron, J. M., Farmer, J. L., & Owens, C. M. (2010). Estimating individual treatment effects from multiple-baseline data: A Monte Carlo study for multilevel-modeling approaches. *Behavior Research Methods, 42*(4), 930–943. https://doi.org/10.3758/BRM.42.4.930

Ferron, J. M., Goldstein, H., Olszewski, A., & Rohrer, L. (2020). Indexing effects in single-case experimental designs by estimating the percent of goal obtained. *Evidence-Based Communication Assessment and Intervention,* 14(1-2), 6–27. https://doi.org/10.1080/17489539.2020.1732024

Ferron, J. M., Hogarty, K. Y., Dedrick, R. F., Hess, M. R., Niles, J. D., & Kromrey, J. D. (2008). Reporting results from multilevel analyses. In A. A. O'Connell and D. B. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 391–426). Information Age Publishing.

Feeney, T. J., & Ylvisaker, M. (2003). Context-sensitive behavioral supports for young children with TBI: short-term effects and long-term outcome. *The Journal of Head Trauma Rehabilitation*, *18*(1), 33–51.

Feeney, T., & Ylvisaker, M. (2006). Context-sensitive cognitive-behavioural supports for young children with TBI: A replication study. *Brain Injury, 20*(6), 629–645. https://doi.org/10.1080/02699050600744194

Feeney, T. J., & Ylvisaker, M. (2008). Context-sensitive cognitive-behavioral supports for young children with TBI: a second replication study. *Journal of Positive Behavior Interventions, 10*(2), 115–128. https://doi.org/10.1177/1098300707312540

Fisher, W. W., Kelley, M. E., & Lomas, J. E. (2003). Visual aids and structured criteria for improving visual inspection and interpretation of single-case designs. *Journal of Applied Behavior Analysis, 36*(3), 387–406. https://doi.org/10.1901/jaba.2003.36-387

Fletcher, D., Boon, R. T., & Cihak, D. F. (2010). Effects of the TOUCHMATH program compared to a number line strategy to teach addition facts to middle school students with moderate intellectual disabilities. *Education and Training in Autism and Developmental Disabilities, 45*(3), 449–458. https://www.jstor.org/stable/23880117?seq=1

Gage, N. A., & Lewis, T. J. (2014). Hierarchical linear modeling meta-analysis of single-subject design research. *The Journal of Special Education, 48*(1), 3–16. https://doi.org/10.1177/0022466912443894

Ganz, J. B., & Ayres, K. M. (2018). Methodological standards in single-case experimental design: Raising the bar. *Research in Developmental Disabilities, 79*(1), 3–9. https://doi.org/10.1016/j.ridd.2018.03.003

Geist, K., & Hitchcock, J. H. (2014). Single case design studies in music therapy: Resurrecting experimental evidence in small group and individual music therapy clinical settings. *Journal of Music Therapy, 51*(4), 293–309. https://doi.org/10.1093/jmt/thu032

Gigerenzer, G. (2004). Mindless statistics. *Journal of Socio-Economics, 33*(5), 587–606. https://doi.org/10.1016/j.socec.2004.09.033

Goldstein, H., Browne, W., & Rasbash, J. (2002). Partitioning variation in multilevel models. *Understanding Statistics, 1*(4), 223-231. https://doi.org/10.1207/S15328031US0104_02

Hagopian, L. P. (2020). The consecutive controlled case series: Design, data-analytics, and reporting methods supporting the study of generality. *Journal of Applied Behavior Analysis*, 53(2), 596–619. https://doi.org/10.1002/jaba.691

Harrington, M., & Velicer, W. F. (2015). Comparing visual and statistical analysis in single-case studies using published studies. *Multivariate Behavioral Research, 50*(2), 162–183. https://doi.org/10.1080/00273171.2014.973989

Hershberger, S. L., Wallace, D. D., Green, S. B., & Marquis, J. G. (1999). Meta-analysis of single-case data. In R. H. Hoyle (Ed.), *Statistical strategies for small sample research* (pp. 107–132). Sage.

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*(1), 51–64. https://doi.org/10.1016/j.jcbs.2013.10.002

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*(2), 165−179. https://doi.org/10.1177/001440290507100203

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). Routledge.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting, 22*(4), 679–688. https://doi.org/10.1016/j.ijforecast.2006.03.001

Jamshidi, L., Heyvaert, M., & Van den Noortgate, W. (2017). *Review of systematic reviews and meta-analyses of single-subject experimental studies*. In The Annual Meeting of the American Education Research Association (AERA), Date: 2017/04/27-2017/05/01, Location: San Antonio (TX), USA. Retrieved on from https://lirias.kuleuven.be/retrieve/513027

Karazsia, B. T. (2018). Editorial: New instructions for single-subject research in the Journal of Pediatric Psychology. *Journal of Pediatric Psychology, 43*(6), 585–587. https://doi.org/10.1093/jpepsy/jsy039

Kavelaars, X. (2020). Going multivariate in clinical trial studies: A Bayesian framework for multiple binary outcomes. In R. van de Schoot & M. Miočević (Eds.), *Small sample size solutions: A guide for applied researchers and practitioners* (pp. 139–154). Routledge.

Kazdin, A. E. (1977). Assessing the clinical or applied importance of behavior change through social validation. *Behavior Modification, 1*(4), 427–452. https://doi.org/10.1177/014544557714001

Kazdin, A. E. (1978). Methodological and interpretive problems of single-case experimental designs. *Journal of Consulting and Clinical Psychology, 46*(4), 629–642. https://doi.org/10.1037/0022-006X.46.4.629

Kazdin, A. E. (2019). Single-case experimental designs. Evaluating interventions in research and clinical practice. *Behaviour Research and Therapy, 117*, 3–17. https://doi.org/10.1016/j.brat.2018.11.015

Kennedy, C. H. (2005). *Single-case designs for educational research*. Pearson.

Klaassen, F. (2020). Combining evidence over multiple individual cases. In R. van de Schoot & M. Miočević (Eds.), *Small sample size solutions: A guide for applied researchers and practitioners* (pp. 126–138). Routledge.

Koegel, L. K., Koegel, R. L., Hurley, C., & Frea, W. D. (1992). Improving social skills and disruptive behavior in children with autism through self-management. *Journal of Applied Behavior Analysis*, *25*(2), 341–353. https://doi.org/10.1901/jaba.1992.25-341

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

Kratochwill, T. R., Levin, J. R., Horner, R. H., & Swoboda, C. M. (2014). Visual analysis of single-case intervention research: Conceptual and methodological issues. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: Methodological and statistical advances* (pp. 91–125). American Psychological Association. https://doi.org/10.1037/14376-004

Lane, J. D., Ledford, J. R., & Gast, D. L. (2017). Single-case experimental design: current standards and applications in occupational therapy. *American Journal of Occupational Therapy*, *71*(2), 7102300010p1–7102300010p9. https://doi.org/10.5014/ajot.2017.022210

Lanovaz, M. J., Turgeon, S., Cardinal, P., & Wheatley, T. L. (2019). Using single-case designs in practical settings: Is within-subject replication always necessary?

Perspectives on Behavior Science, 42(1), 153–162. https://doi.org/10.1007/s40614-018-0138-9

Laski, K. E., Charlop, M. H., & Schreibman, L. (1988). Training parents to use the natural language paradigm to increase their autistic children's speech. *Journal of Applied Behavior Analysis, 21*(4), 391–400. https://doi.org/10.1901/jaba.1988.21-391

Ledford, J. R. (2018). No randomization? No problem: Experimental control and random assignment in single case research. *American Journal of Evaluation, 39*(1), 71–90. https://doi.org/10.1177/1098214017723110

Ledford, J. R., Barton, E. E., Hardy, J. K., Elam, K., Seabolt, J., Shanks, M., ... & Kaiser, A. (2016). What equivocal data from single case comparison studies reveal about evidence-based practices in early childhood special education. *Journal of Early Intervention, 38*(2), 79–91. https://doi.org/10.1177/1053815116648000

Ledford, J. R., Barton, E. E., Severini, K. E., & Zimmerman, K. N. (2019). A primer on single-case research designs: Contemporary use and analysis. *American Journal on Intellectual and Developmental Disabilities, 124*(1), 35–56. https://doi.org/10.1352/1944-7558-124.1.35

Ledford, J. R., & Gast, D. L. (2018). Combination and other designs. In D. L. Gast & J. R. Ledford (Eds.), *Single case research methodology: Applications in special education and behavioral sciences* (3rd ed.) (pp. 335–364). Routledge.

Leek, J. T., & Jager, L. R. (2017). Is most published research really false? *Annual Review of Statistics & Its Application, 4*, 109–122. https://doi.org/10.1146/annurev-statistics-060116-054104.

Levin, J. R., Ferron, J. M., & Gafurov, B. S. (2017). Additional comparisons of randomization-test procedures for single-case multiple-baseline designs: Alternative effect types. *Journal of School Psychology, 63,* 13–34. https://doi.org/10.1016/j.jsp.2017.02.003

Lloyd, B. P., Finley, C. I., & Weaver, E. S. (2018). Experimental analysis of stereotypy with applications of nonparametric statistical tests for alternating treatments designs. *Developmental Neurorehabilitation, 21*(4), 212–222. https://doi.org/10.3109/17518423.2015.1091043

Lobo, M. A., Moeyaert, M., Cunha, A. B., & Babik, I. (2017). Single-case design, analysis, and quality assessment for intervention research. *Journal of Neurologic Physical Therapy, 41*(3), 187–197. https://doi.org/10.1097/NPT.0000000000000187

L'Abbé, K. A., Detsky, A. S., & O'Rourke, K. (1987). Meta-analysis in clinical research. *Annals of Internal Medicine, 107*(2), 224−233. https://doi.org/10.7326/0003-4819-107-2-224

Maggin, D. M., Briesch, A. M., & Chafouleas, S. M. (2013). An application of the What Works Clearinghouse standards for evaluating single-subject research: Synthesis of the self-management literature base. *Remedial and Special Education, 34*(1), 44−58. https://doi.org/10.1177/0741932511435176

Maggin, D. M., Briesch, A. M., Chafouleas, S. M., Ferguson, T. D., & Clark, C. (2014). A comparison of rubrics for identifying empirically supported practices with single-case research. *Journal of Behavioral Education, 2*3(2), 287–311. https://doi.org/10.1007/s10864-013-9187-z

Maggin, D. M., Cook, B. G., & Cook, L. (2018). Using single-case research designs to examine the effects of interventions in special education. *Learning Disabilities Research & Practice, 33*(4), 182–191. https://doi.org/10.1111/ldrp.12184

Manolov, R. (2018). Linear trend in single-case visual and quantitative analyses. *Behavior Modification, 42*(5), 684–706. https://doi.org/10.1177/0145445517726301

Manolov, R., & Moeyaert, M. (2017). How can single-case data be analyzed? Software resources, tutorial, and reflections on analysis. *Behavior Modification, 41*(2), 179–228. https://doi.org/10.1177/0145445516664307

Manolov, R., Tanious, R., De, T. K., & Onghena, P. (2020). Assessing consistency in single-case alternation designs. *Behavior Modification*. Advance online publication. https://doi.org/10.1177/0145445520923990

McDougall, D., Hawkins, J., Brady, M., & Jenkins, A. (2006). Recent innovations in the changing criterion design: Implications for research and practice in special education. *The Journal of Special Education, 40*(1), 2–15. https://doi.org/10.1177/00224669060400010101

Mengersen, K., McGree, J. M., & Schmid, C. H. (2015). Statistical analysis of N-of-1 trials. In J. Nikles & G. Mitchell (Eds.), *The essential guide to N-of-1 trials in health* (pp.135-153). Springer.

Merlo, J., Chaix, B., Yang, M., Lynch, J., & Råstam, L. (2005). A brief conceptual tutorial of multilevel analysis in social epidemiology: linking the statistical concept of clustering to the idea of contextual phenomenon. *Journal of Epidemiology & Community Health, 59*(6), 443-449. http://dx.doi.org/10.1136/jech.2004.023473

Michiels, B., & Onghena, P. (2019). Randomized single-case AB phase designs: Prospects and pitfalls. *Behavior Research Methods, 51*(6), 2454–2476. https://doi.org/10.3758/s13428-018-1084-x

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

Moeyaert, M., Rindskopf, D., Onghena, P., & Van den Noortgate, W. (2017). Multilevel modeling of single-case data: A comparison of maximum likelihood and Bayesian estimation. *Psychological Methods, 22*(4), 760–778. https://doi.org/10.1037/met0000136

Natesan, P. (2019). Fitting Bayesian models for single-case experimental designs. *Methodology*, *15*(4), 147−156. https://doi.org/10.1027/1614-2241/a000180

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*(2), 241–301. https://doi.org/10.1037/1082-989X.5.2.241

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

Normand, M. P. (2016). Less is more: Psychologists can learn more by studying fewer people. *Frontiers in Psychology, 7,* e934. https://doi.org/10.3389/fpsyg.2016.00934

Olive, M. L., & Smith, B. W. (2005). Effect size calculations and single subject designs. *Educational Psychology, 25*(2-3), 313–324. https://doi.org/10.1080/0144341042000301238

Onghena, P., & Edgington, E. S. (1994). Randomization tests for restricted alternating treatments designs. *Behaviour Research and Therapy, 32*(7), 783–786. https://doi.org/10.1016/0005-7967(94)90036-1

Onghena, P., & Edgington, E. S. (2005). Customization of pain treatments: Single-case design and analysis. *Clinical Journal of Pain, 21*(1), 56–68. https://doi.org/10.1097/00002508-200501000-00007

Parker, R. I., Cryer, J., & Byrns, G. (2006). Controlling baseline trend in single-case research. *School Psychology Quarterly, 21*(4), 418–443. https://doi.org/10.1037/h0084131

Parker, R. I., & Vannest, K. J. (2009). An improved effect size for single-case research: Nonoverlap of all pairs. *Behavior Therapy, 40*(4), 357−367. https://doi.org/10.1016/j.beth.2008.10.006

Parker, R. I., & Vannest, K. J. (2012). Bottom-up analysis of single-case research designs. *Journal of Behavioral Education, 21*(3), 254−265. https://doi.org/10.1007/s10864-012-9153-1

Peugh, J. L. (2010). A practical guide to multilevel modeling. *Journal of School Psychology, 48*(1), 85-112. https://doi.org/10.1016/j.jsp.2009.09.002

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & R Core Team. (2020). *nlme: Linear and nonlinear mixed effects models*. R package version 3.1-148, https://CRAN.R-project.org/package=nlme

Plavnick, J. B., & Ferreri, S. J. (2013). Single-case experimental designs in educational research: A methodology for causal analyses in teaching and learning. *Educational Psychology Review, 25*(4), 549−569. https://doi.org/10.1007/s10648-013-9230-6

Radley, K. C., Dart, E. H., Fischer, A. J., & Collins, T. A. (2020). Publication trends for single-case methodology in school psychology: A systematic review. *Psychology in the Schools, 57*(5), 683−698. https://doi.org/10.1002/pits.22359

Romeiser-Logan, L., Slaughter, R., & Hickman, R. (2017). Single-subject research designs in pediatric rehabilitation: A valuable step towards knowledge translation. *Developmental Medicine & Child Neurology, 59*(6), 574−580. https://doi.org/10.1111/dmcn.13405

Scruggs, T. E., & Mastropieri, M. A. (1998). Summarizing single-subject research: Issues and applications. *Behavior Modification, 22*(3), 221−242. https://doi.org/10.1177/01454455980223001

Shadish, W. R., Hedges, L. V., & Pustejovsky, J. E. (2014). Analysis and meta-analysis of single-case designs with a standardized mean difference statistic: A primer and applications. *Journal of School Psychology, 52*(2), 123–147. https://doi.org/10.1016/j.jsp.2013.11.005

Shadish, W. R., Kyse, E. N., & Rindskopf, D. M. (2013). Analyzing data from single-case designs using multilevel models: New applications and some agenda items for future research. *Psychological Methods, 18*(3), 385–405. https://doi.org/10.1037/a0032964

Shadish, W. R., & Sullivan, K. J. (2011). Characteristics of single-case designs used to assess intervention effects in 2008. *Behavior Research Methods, 43*(4), 971−980. https://doi.org/10.3758/s13428-011-0111-y

Sharp, S. J., Thompson, S. G., & Altman, D. G. (1996). The relation between treatment benefit and underlying risk in meta-analysis. *BMJ, 313*(7059), 735-738. https://doi.org/10.1136/bmj.313.7059.735

Song, F. (1999). Exploring heterogeneity in meta-analysis: Is the L'Abbe plot useful?. *Journal of Clinical Epidemiology, 52*(8), 725-730. https://doi.org/10.1016/S0895-4356(99)00066-9

Tanious, R., De, T. K., Michiels, B., Van den Noortgate, W., & Onghena, P. (2020). Assessing consistency in single-case A-B-A-B phase designs. *Behavior Modification*, *44*(4), 518-551. https://doi.org/10.1177/0145445519837726

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

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*, 103414. https://doi.org/10.1016/j.brat.2019.103414

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

Tate, R. L., & Perdices, M. (2019). *Single-case experimental designs for clinical research and neurorehabilitation settings: Planning, conduct, analysis, and reporting*. Routledge.

Tate, R. L., Perdices, M., Rosenkoetter, U., McDonald, S., Togher, L., …, Wilson, B. (2016). The Single-Case Reporting guideline In BEhavioural interventions (SCRIBE) 2016 statement. *Journal of School Psychology, 56*, 133−142. https://doi.org/10.1016/j.jsp.2016.04.001

Vannest, K. J., Parker, R. I., Davis, J. L., Soares, D. A., & Smith, S. L. (2012). The Theil–Sen slope for high-stakes decisions from progress monitoring. *Behavioral Disorders, 37*(4), 271–280. https://doi.org/10.1177/019874291203700406

Wolery, M., Gast, D. L., & Ledford, J. R. (2018). Comparative designs. In D. L. Gast & J. R. Ledford (Eds.), *Single case research methodology: Applications in special education and behavioral sciences* (3rd ed.) (pp. 283–334). Routledge.

Table 1

Descriptive summary of several multiple-baseline data sets

Data	Tier	Phase A Mean	Phase B Mean	Phase A SD	Phase B SD	Phase A Trend	Phase B Trend
Allen	Tier 1	26.94	78.46	7.92	9.83	2.03	0.07
	Tier 2	36.43	72.06	10.58	9.74	2.68	0.75
	Tier 3	37.46	74.36	12.07	11.58	1.24	0.67
Laski echolalic	Tier 1	39.81	82.35	13.90	5.56	-3.24	-0.18
	Tier 2	58.21	83.16	5.17	5.36	0.82	0.97
	Tier 3	68.28	73.25	8.10	11.53	2.28	1.23
	Tier 4	54.88	76.88	15.03	6.20	-1.56	-0.34
Laski nonverbal	Tier 1	30.66	59.70	8.40	10.62	5.88	-0.55
	Tier 2	34.62	78.62	9.52	8.57	-0.43	-0.77
	Tier 3	0.00	49.41	0.00	16.18	0.00	4.22
	Tier 4	0.00	47.76	0.00	14.46	0.00	2.27
	Tier 5	20.27	59.57	13.37	12.92	0.98	-3.02

Table 2

Results of applying a two-level model to each data set

Data	Tier	Initial baseline level	Baseline trend	Immediate change in level	Change in trend
Allen	Average	29.02	0.76	31.72	-0.22
	SD	1.60	0.38	14.28	0.00
	Tier 1	27.32	0.35	47.00	-0.22
	Tier 2	29.87	0.96	24.09	-0.22
	Tier 3	29.88	0.96	24.08	-0.22
Laski echolalic	Average	58.47	-0.71	25.29	0.89
	SD	3.35	1.96	25.27	1.49
	Tier 1	54.41	-3.03	55.21	2.65
	Tier 2	59.40	-0.55	23.39	0.74
	Tier 3	61.30	1.55	-4.05	-0.73
	Tier 4	58.79	-0.81	26.61	0.91
Laski nonverbal	Average	15.39	0.37	38.69	0.22
	SD	15.15	0.25	2.93	2.27
	Tier 1	23.24	0.52	39.67	-1.25
	Tier 2	34.49	0.50	42.58	-2.01
	Tier 3	1.07	0.15	36.45	3.57
	Tier 4	0.50	0.20	35.99	2.27
	Tier 5	17.65	0.45	38.74	-0.47

Note. Average corresponds to the fixed effect estimate. SD denotes the standard deviation and corresponds to the random effect estimates. Tiers – empirical Bayes estimates for each tier. Initial baseline level corresponds to the intercept.

BIOGRAPHICAL NOTE

Rumen Manolov, PhD, is an associate professor at the Faculty of Psychology at the University of Barcelona. His investigation is focused on single-case designs, specifically, testing and proposing data analytical techniques, as well as developing and promoting R code and user-friendly Shiny applications for single-case data analysis.

René Tanious is a doctoral candidate at the Faculty of Psychology and Educational Sciences at KU Leuven, Belgium. His current research interests include single-case experimental designs, development of effect size measures, and combining statistical and visual analysis for single-case experimental designs.

Figure 1

A modified Brinley plot for an ABAB design, replicated across six participants. Data gathered by Feeney and Ylvisaker (2003, 2006, 2008). The points represent phase means. The solid diagonal line represents equality between conditions: the points below it indicate a reduction (here, an improvement) after the intervention. The dotted horizontal line represents a supposed desired level of behavior after the intervention: the points below it represent achieving this desired level.

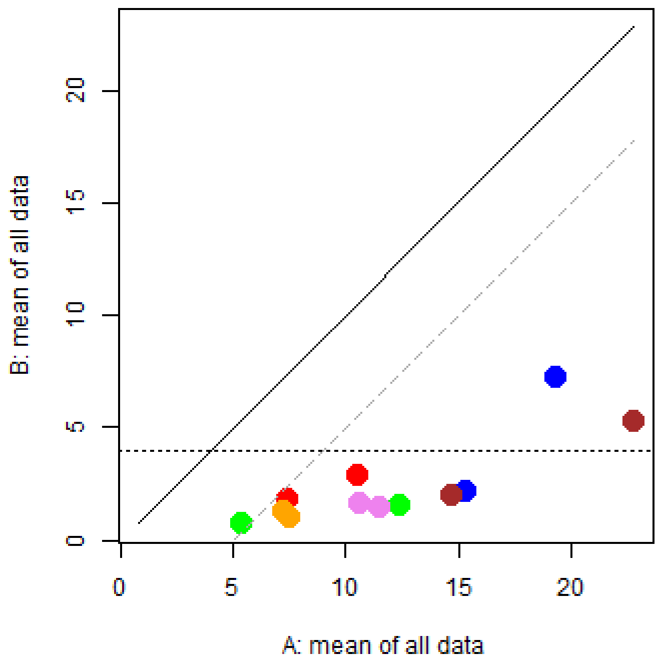

Figure 2

Fictitious data from Example 1

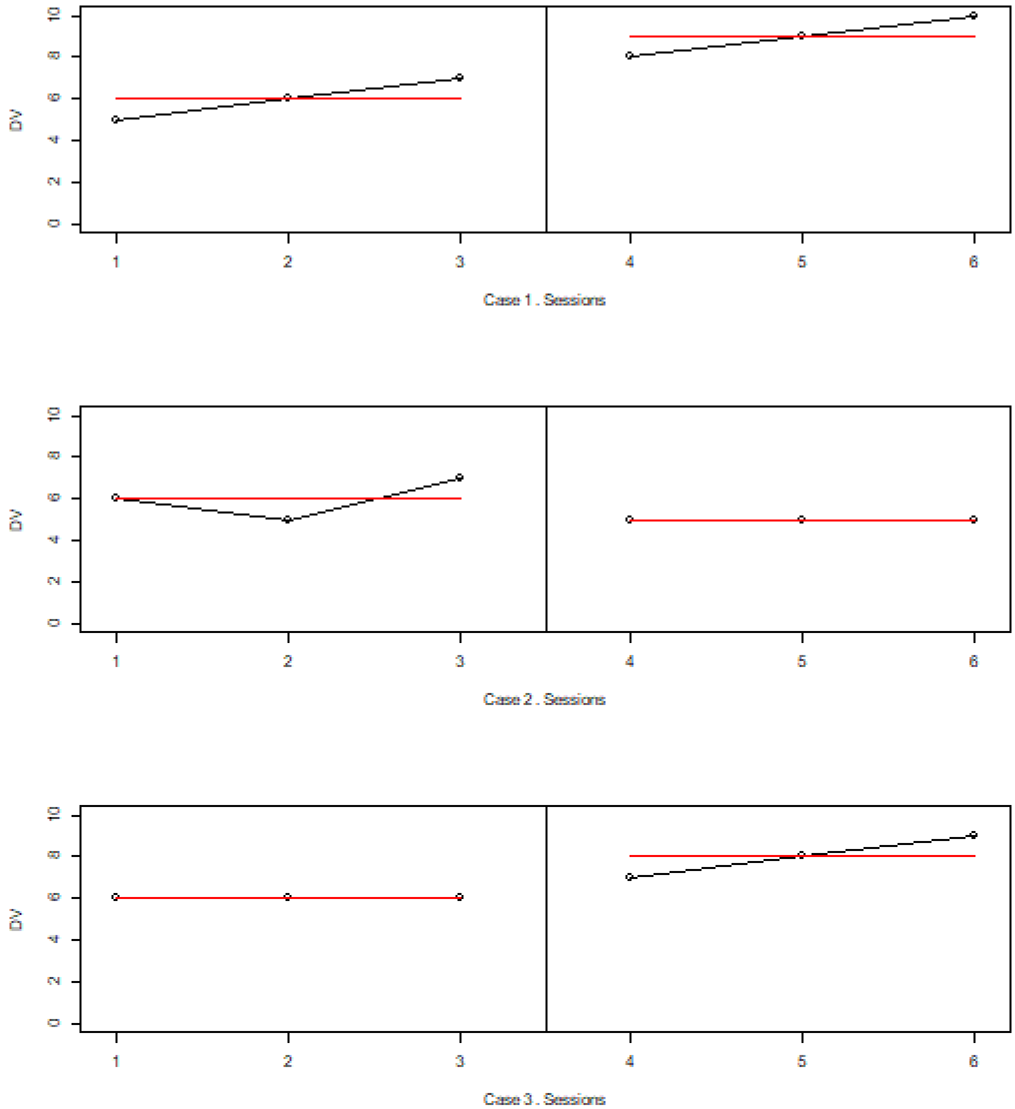

Figure 3

For the data from Example 2. Upper left panel: the closer the dots horizontally the greater the consistency in condition A. Upper right panel: the closer the dots vertically the greater the consistency in mean level in condition B. Lower panel: the closer the dots to the dashed diagonal line, the greater the consistency of effect (difference in means).

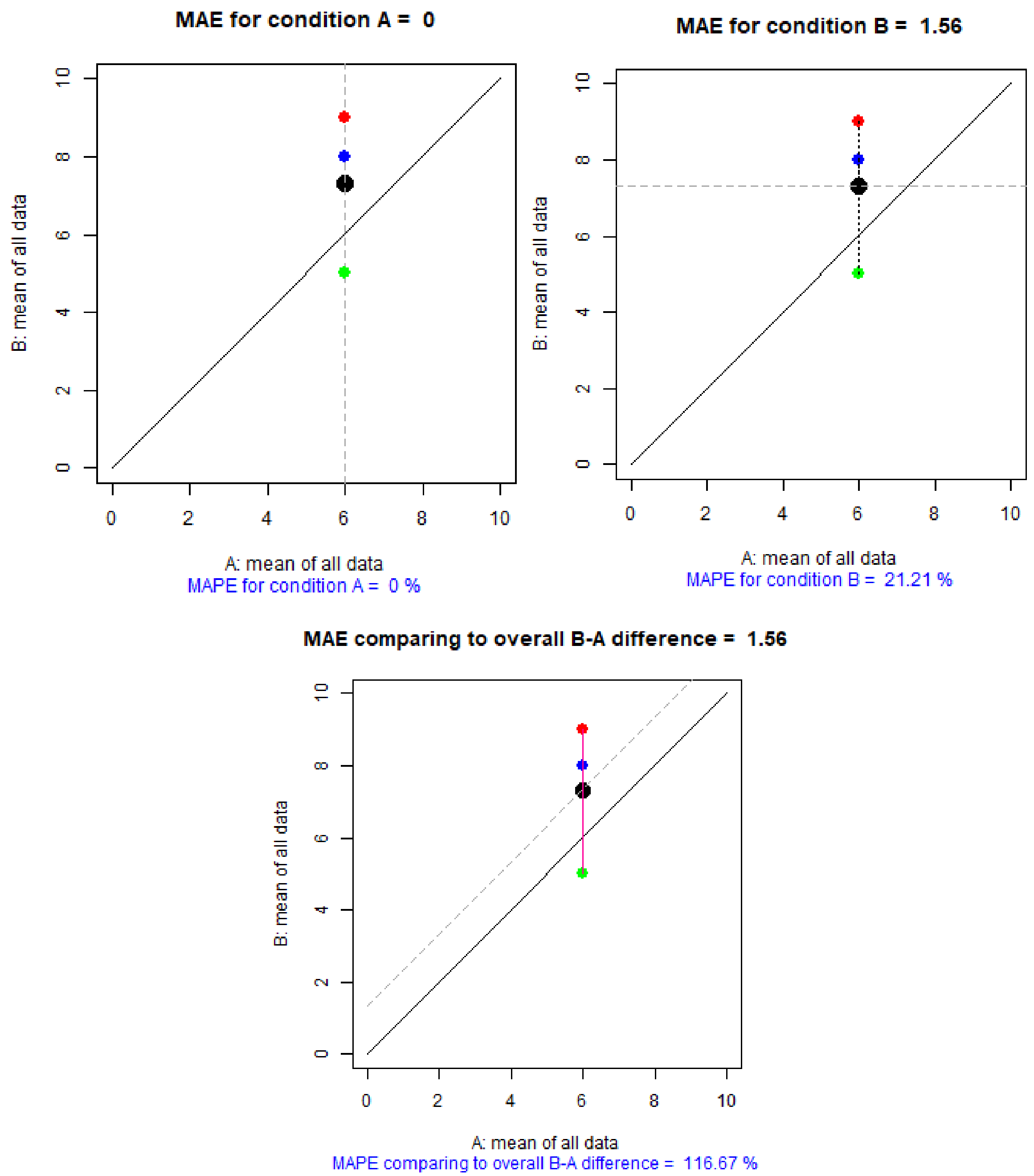

Figure 4

Fictitious data from Example 2

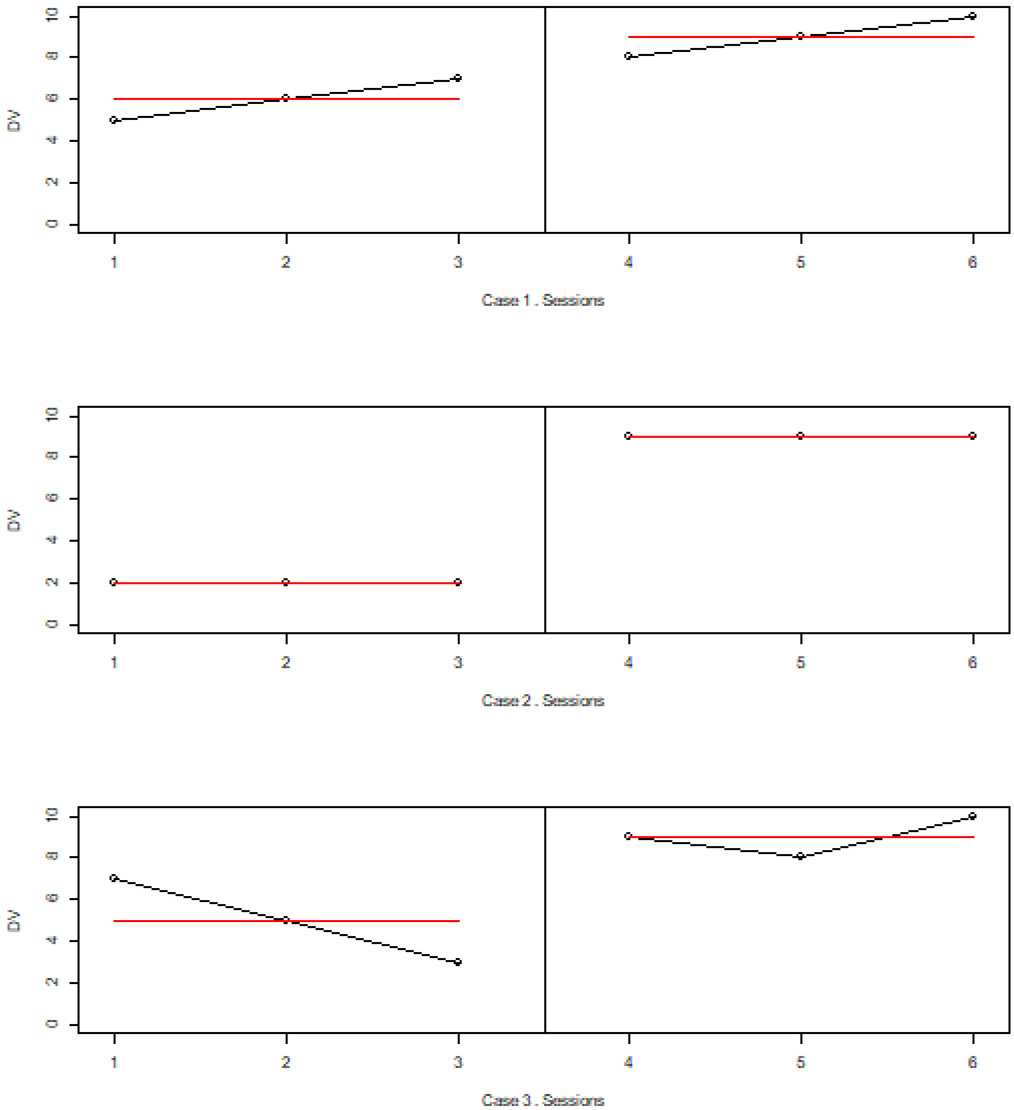

Figure 5

For the data from Example 2. Upper left panel: the closer the dots horizontally the greater the consistency in condition A. Upper right panel: the closer the dots vertically the greater the consistency in mean level in condition B. Lower panel: the closer the dots to the dashed diagonal line, the greater the consistency of effect (difference in means).

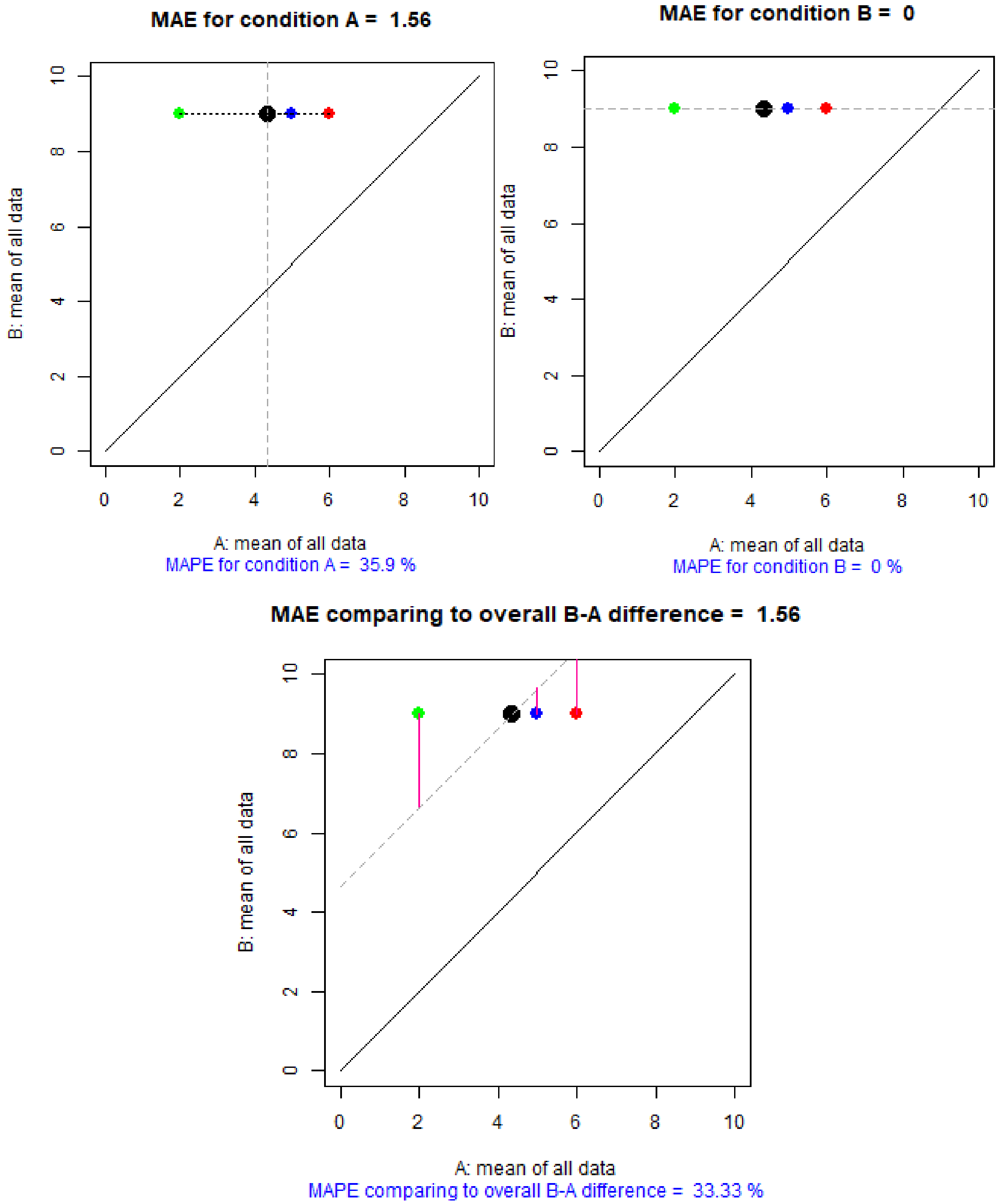

Figure 6

Fictitious data from Example 3

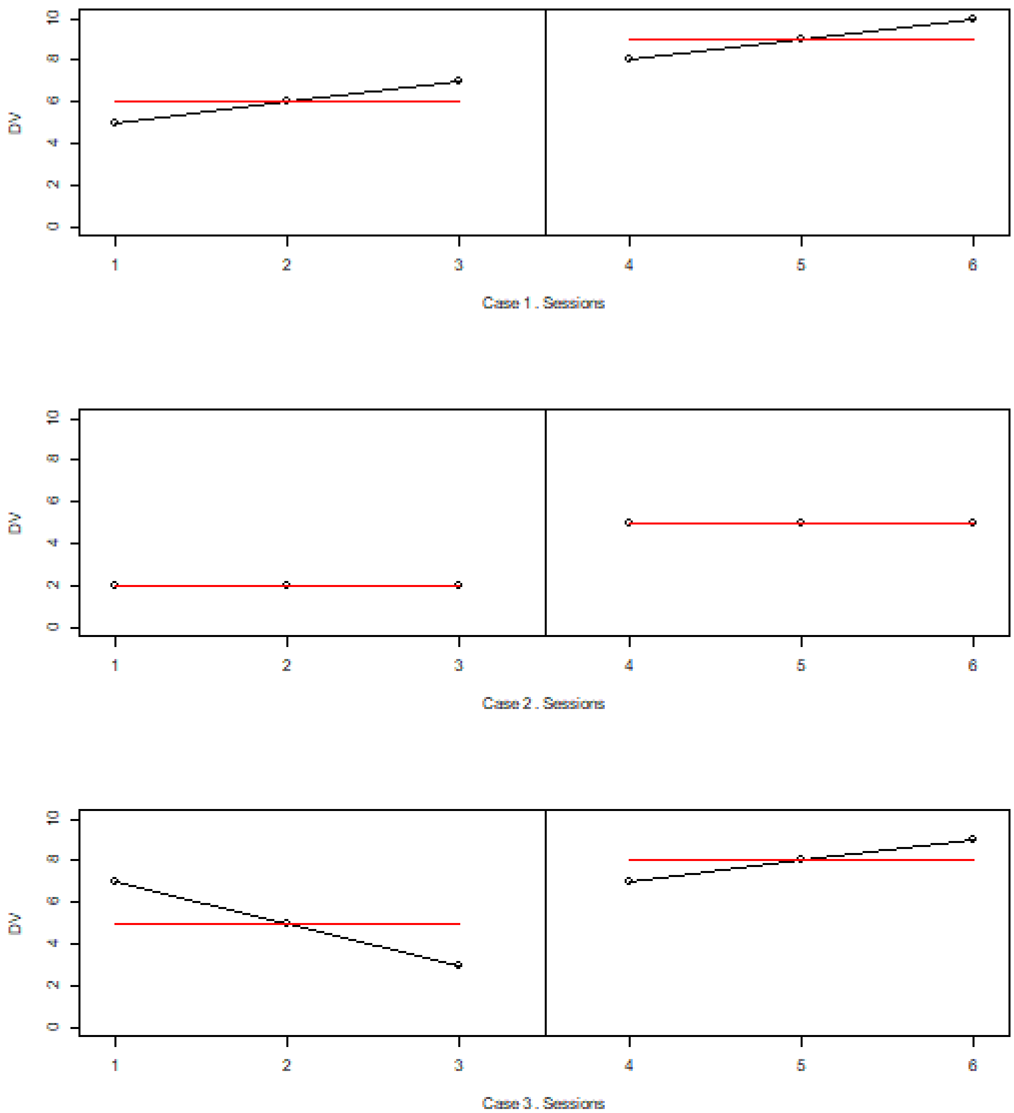

Figure 7

For the data from Example 3. Upper left panel: the closer the dots horizontally the greater the consistency in condition A. Upper right panel: the closer the dots vertically the greater the consistency in mean level in condition B. Lower panel: the closer the dots to the dashed diagonal line, the greater the consistency of effect (difference in means).

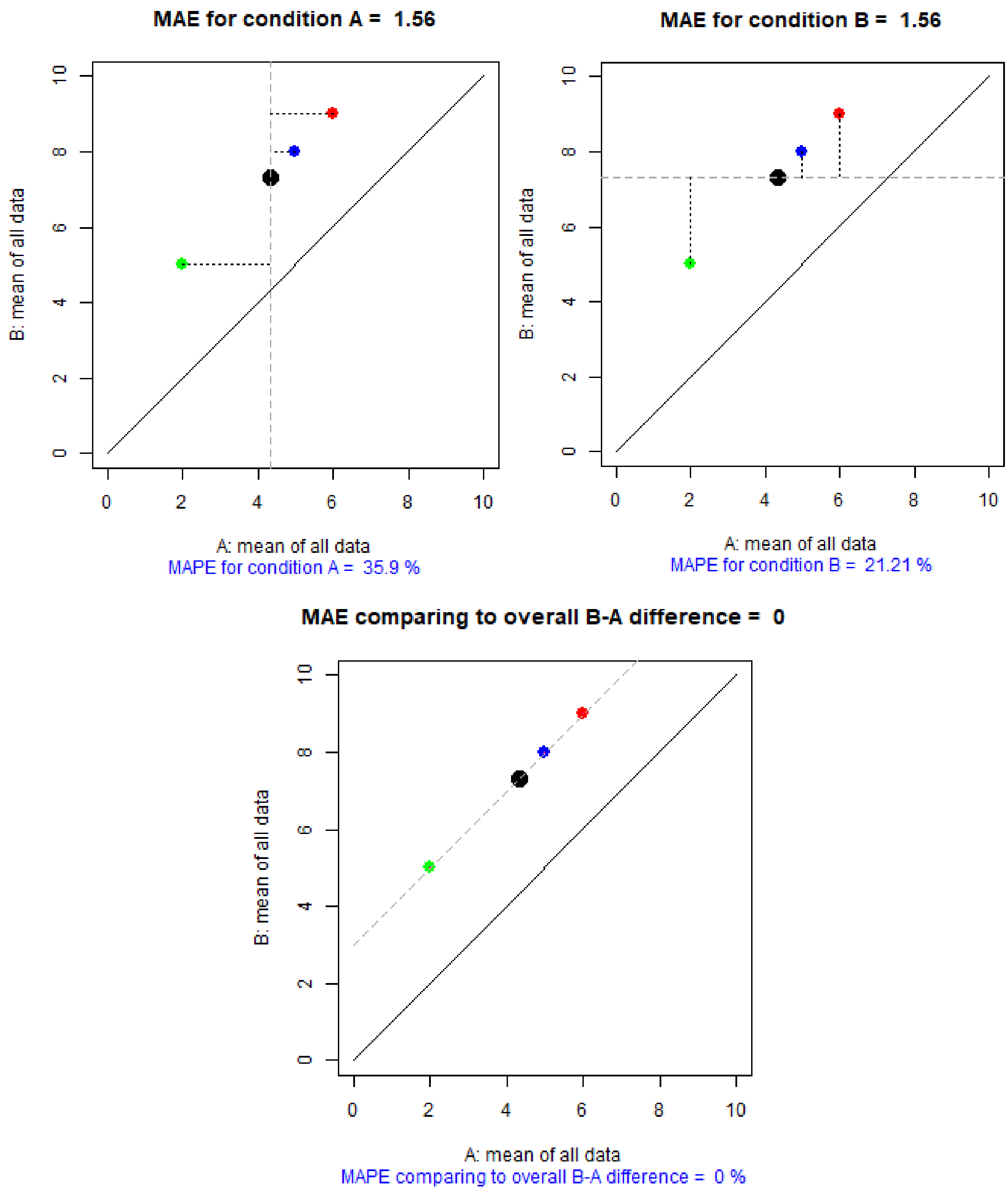

Figure 8

Data on the percentage of correct responding per session by a participant called Holly, diagnosed with Autism Spectrum Disorder and Intellectual Developmental Disorder. The tiers correspond to three behaviors: requesting help, checking out, and ordering food. Data gathered by Allen et al. (2015). The output is from the https://manolov.shinyapps.io/Brinley/ *website. The red horizontal line represents the within-phase mean, which is the basis of the Brinley plot.*

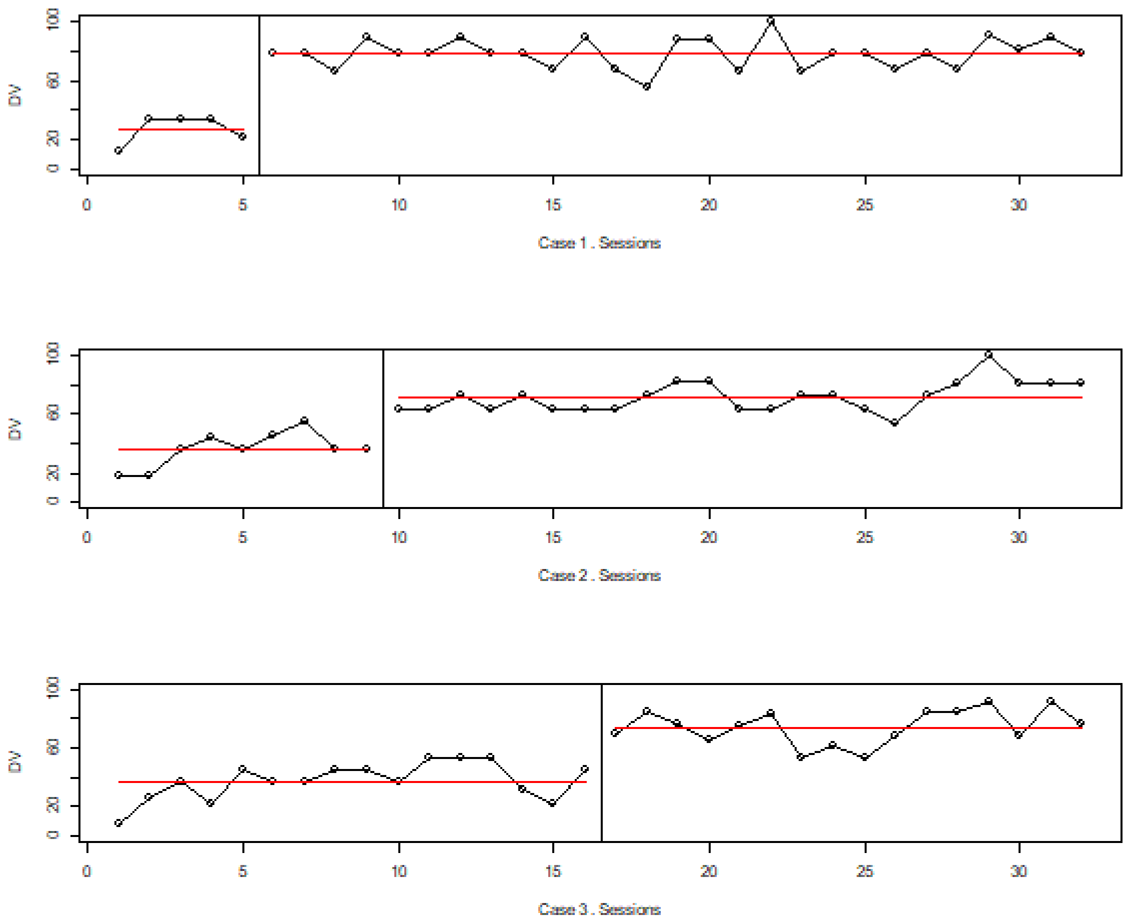

Figure 9

Consistency of level in similar phases, for the Allen et al. (2015) data. The vertical dashed line on the left panel represents the overall mean for condition A: the closer the points horizontally the greater the consistency in this condition. The horizontal dashed line on the right panel represents the overall mean for condition B: the closer the points vertically the greater the consistency in this condition

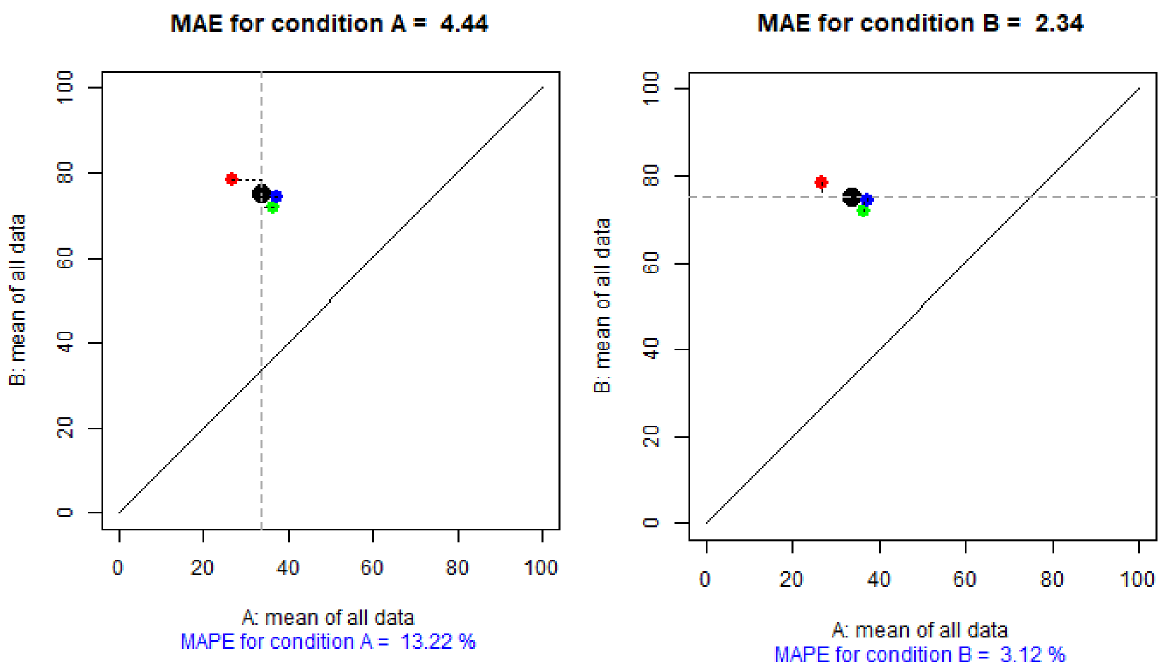

Figure 10

Consistency of effect, for the Allen et al. (2015) data. The diagonal dashed line represents the average difference between conditions A and B. The closer the points vertically to this dashed diagonal line, the greater the consistency of effect (mean difference).

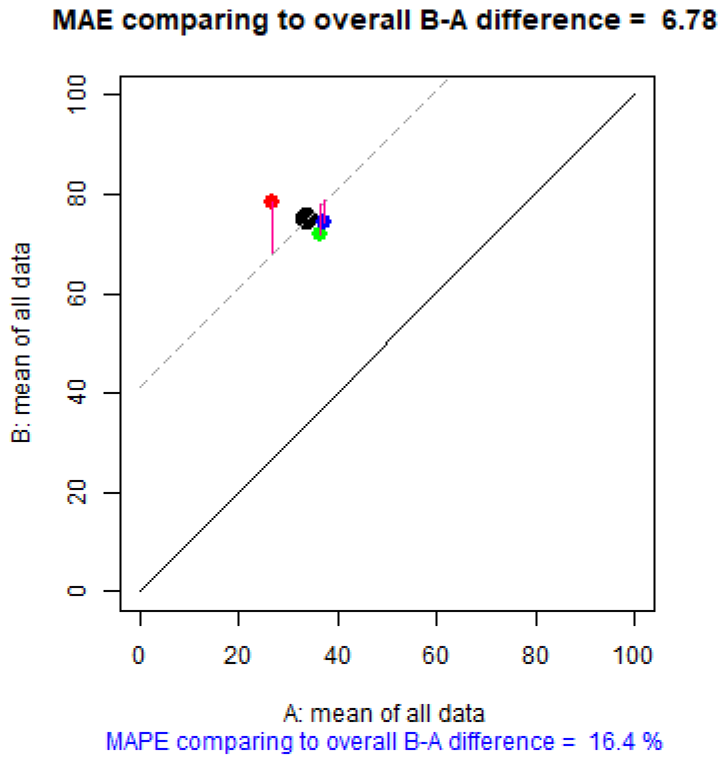

Figure 11

Data on the verbalizations by children with echolalia in free play context, as gathered by Laski et al. (1988). The output is from the https://manolov.shinyapps.io/Brinley/ website. The red horizontal line represents the within-phase mean, which is the basis of the Brinley plot.

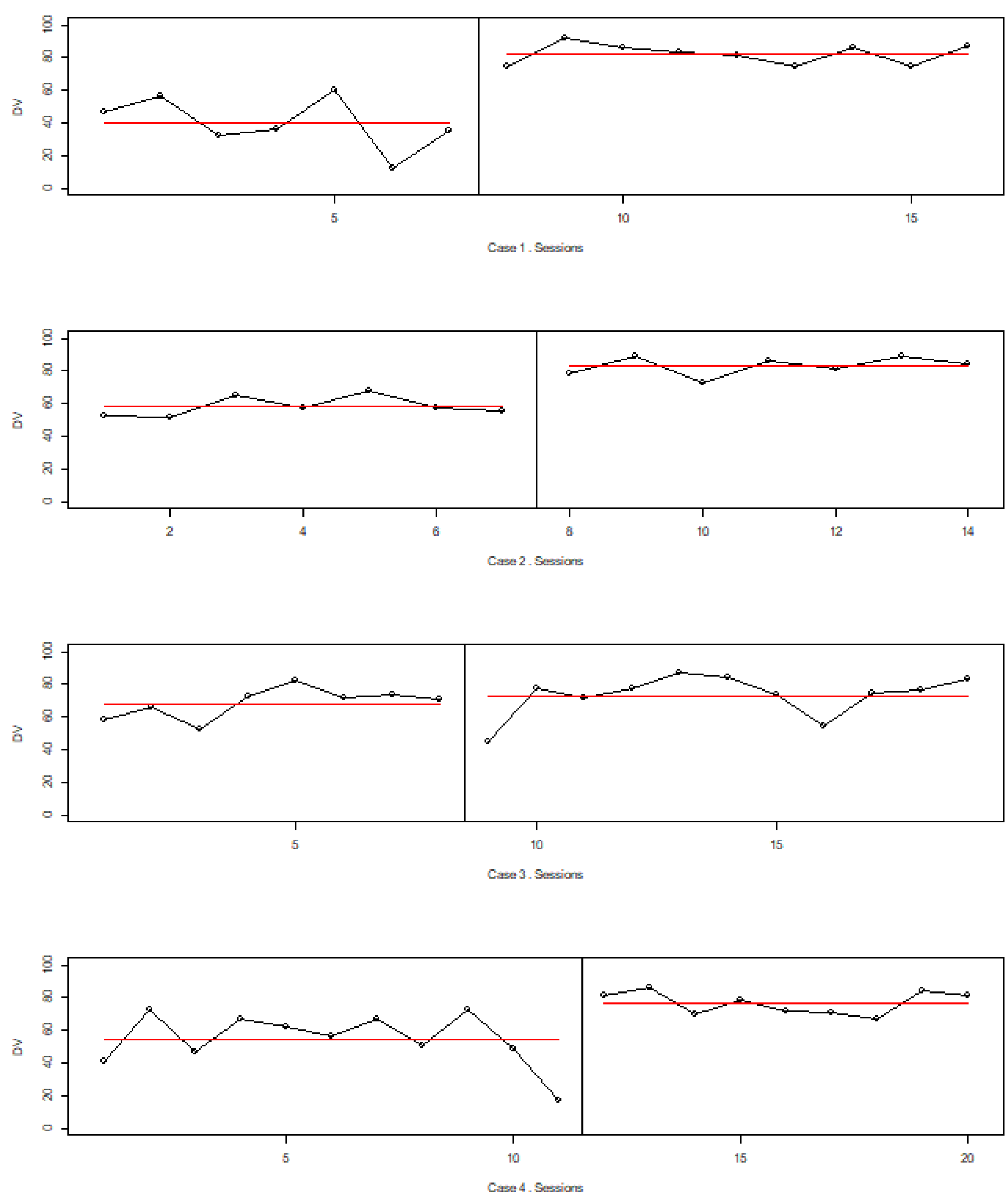

Figure 12

Consistency of level in similar phases, for the Laski et al. (1988) data for the echolalic children. The vertical dashed line on the left panel represents the overall mean for condition A: the closer the points horizontally the greater the consistency in this condition. The horizontal dashed line on the right panel represents the overall mean for condition B: the closer the points vertically the greater the consistency in this condition.

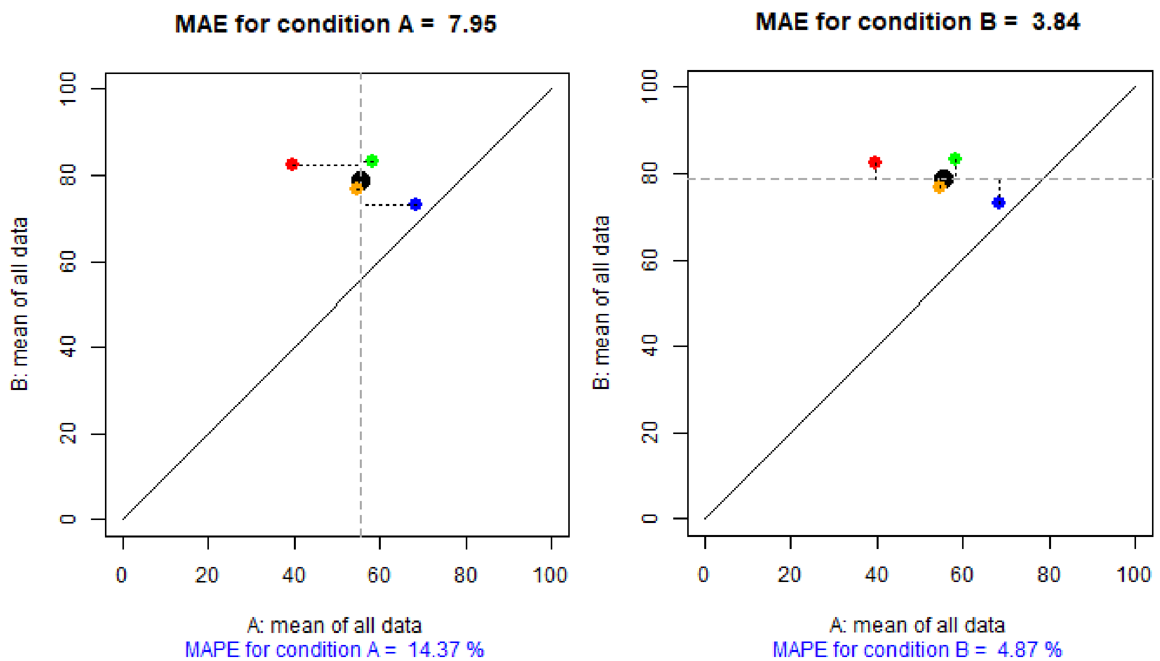

Figure 13

Consistency of effect, for the Laski et al. (1988) data for echolalic children. The diagonal dashed line represents the average difference between conditions A and B. The closer the points vertically to this dashed diagonal line, the greater the consistency of effect (mean difference).

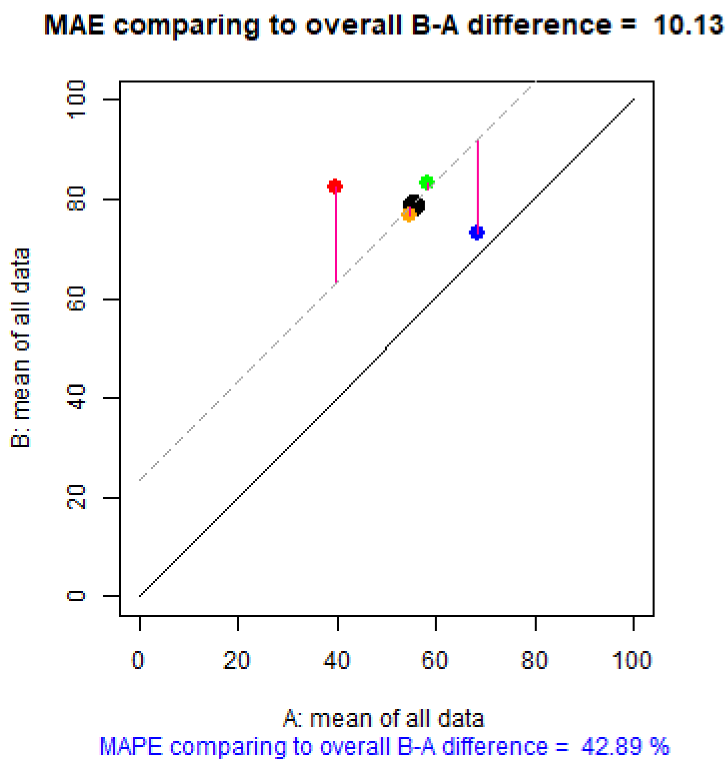

Figure 14

Data on the verbalizations by nonverbal children in free play context, as gathered by Laski et al. (1988). The output is from the https://manolov.shinyapps.io/Brinley/ website. The red horizontal line represents the within-phase mean, which is the basis of the Brinley plot.

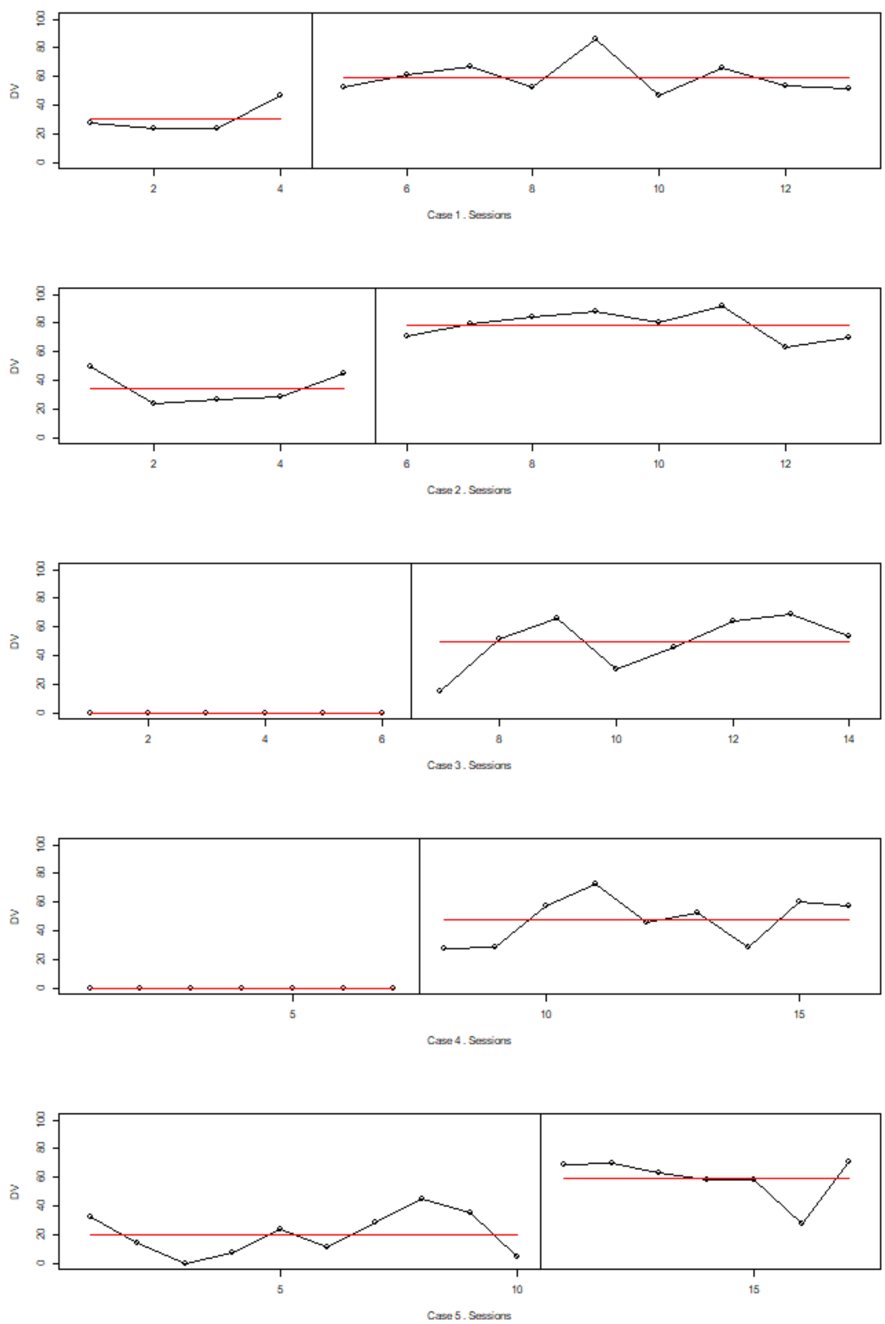

Figure 15

Consistency of level in similar phases, for the Laski et al. (1988) data for the nonverbal children. The vertical dashed line on the left panel represents the overall mean for condition A: the closer the points horizontally the greater the consistency in this condition. The horizontal dashed line on the right panel represents the overall mean for condition B: the closer the points vertically the greater the consistency in this condition.

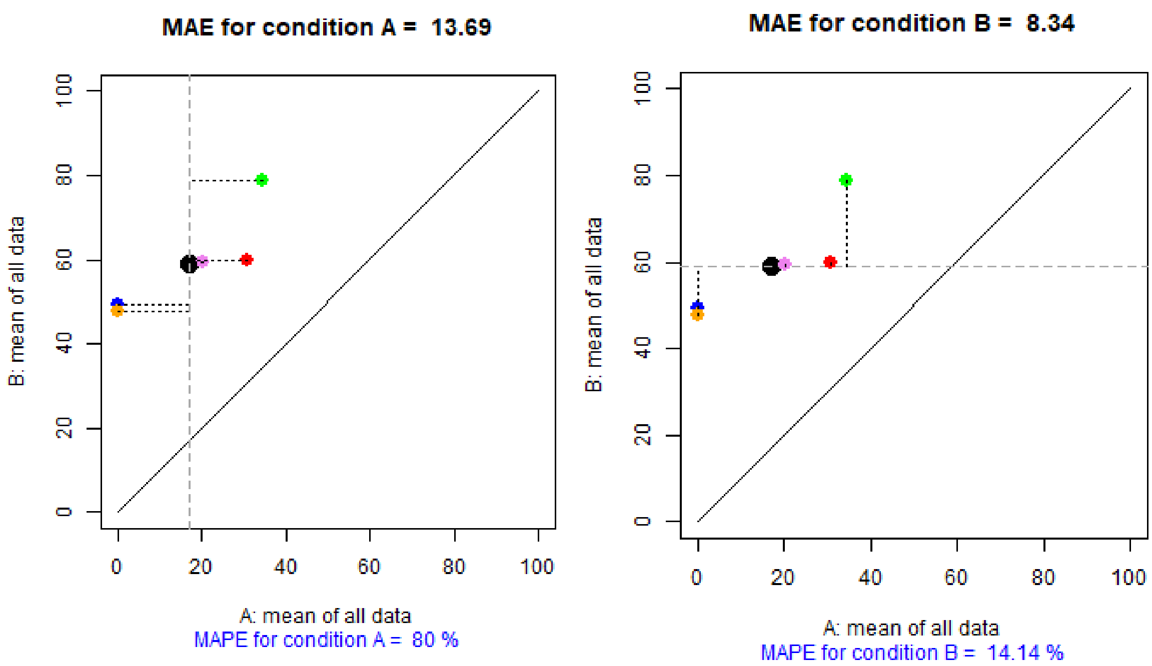

Figure 16

Consistency of effect, for the Laski et al. (1988) data for nonverbal children. The diagonal dashed line represents the average difference between conditions A and B. The closer the points vertically to this dashed diagonal line, the greater the consistency of effect (mean difference).