Analyzing Two-Phase Single-Case Data with Non-overlap and Mean Difference Indices: Illustration, Software Tools, and Alternatives

Rumen Manolov[1]*, José L. Losada[1], Salvador Chacón-Moscoso[2,3]* and Susana Sanduvete-Chaves[2]

[1] *Departamento de Metodología de las Ciencias del Comportamiento, Facultad de Psicología, Universidad de Barcelona, Barcelona, Spain,* [2] *Psicología Experimental, Universidad de Sevilla, Seville, Spain,* [3] *Universidad Autónoma de Chile, Santiago, Chile*

Two-phase single-case designs, including baseline evaluation followed by an intervention, represent the most clinically straightforward option for combining professional practice and research. However, unless they are part of a multiple-baseline schedule, such designs do not allow demonstrating a causal relation between the intervention and the behavior. Although the statistical options reviewed here cannot help overcoming this methodological limitation, we aim to make practitioners and applied researchers aware of the available appropriate options for extracting maximum information from the data. In the current paper, we suggest that the evaluation of behavioral change should include visual and quantitative analyses, complementing the substantive criteria regarding the practical importance of the behavioral change. Specifically, we emphasize the need to use structured criteria for visual analysis, such as the ones summarized in the What Works Clearinghouse *Standards*, especially if such criteria are complemented by visual aids, as illustrated here. For quantitative analysis, we focus on the non-overlap of all pairs and the slope and level change procedure, as they offer straightforward information and have shown reasonable performance. An illustration is provided of the use of these three pieces of information: visual, quantitative, and substantive. To make the use of visual and quantitative analysis feasible, open source software is referred to and demonstrated. In order to provide practitioners and applied researchers with a more complete guide, several analytical alternatives are commented on pointing out the situations (aims, data patterns) for which these are potentially useful.

Keywords: non-experimental, single-case, data analysis, guidelines, methodological quality

INTRODUCTION

The evidence-based practices movement aims to provide guidelines for carrying out methodologically sound research in fields such as psychology (Apa Presidential Task Force on Evidence-Based Practice, 2006) and special education (Odom et al., 2005). According to this movement, the studies providing solid evidence need to meet a series of criteria related to how

an experimental effect is documented and how generality can be established (Maggin et al., 2014). The first of these aspects refers, among other features of the study, to its design and analysis. In the current work, we focus on two-phase designs that do not meet the criteria established by the What Works Clearinghouse *Standards* (Kratochwill et al., 2010), unless they are part of a within-study replication, as in a multiple-baseline design. Two-phase designs may be weaker, from the perspective of internal validity, but they are still used (e.g., Cordery et al., 2010; O'Neill et al., 2013; Finn and McDonald, 2014; Winkens et al., 2014) and can be useful as pilot studies and also due to the fact that establishing the evidence basis of interventions is related to the replication of results and their integration via systematic reviews and meta-analyses (Jenson et al., 2007). Such reviews can offer a comprehensive summary of findings while trying to avoid publication bias, which would take place when excluding studies on the basis of the design. In that sense, it is potentially useful to report the results of all studies and, afterward, consider whether some studies show no differences or negative results (Kratochwill et al., 2001) or whether there are differences according to the design used or the methodological quality of the study. Actually, Gage and Lewis (2014) suggest that experimental control can be used as a moderator variable in meta-analyses.

In this context, the present paper arises from our conviction that practitioners' professional practice, mainly aimed to help individual clients, can also contribute to informing fellow professionals about the results of applying certain interventions. In order to make this contribution possible and in order to be able to translate practice into research certain design and analysis considerations are necessary. The current paper mainly aims to answer two specific questions "What can be done to improve the data analysis in my practice so that its results are more useful to the discipline, despite using a sub-optimal design?" and "How can I easily implement some appropriate analytical techniques?" However, design and data analysis should be considered jointly (Brossart et al., 2014) and this is why we first review some aspects related to how the study is conducted.

Regarding the ways in which a study can be considered as providing evidence, a design implemented as a randomized controlled trial is one option, but it is not always feasible. Another alternative is single-case designs, also referred to as N-of-1 trials (Howick et al., 2011). For this latter option, there are several guidelines on how the studies should be carried out (see Smith, 2012; Maggin et al., 2014, for a review). Two of these guidelines are What Works Clearinghouse *Standards* (Kratochwill et al., 2010) and the Risk of Bias in N-of-1 Trials (RoBiNT) scale by Tate et al. (2013). In brief, the optimal features of a single-case study contributing solid evidence are: to use a design allowing for at least three comparisons between conditions (as in multiple baseline, alternating treatments, and ABAB designs; Barlow et al., 2009); to include randomization in the design when assigning measurement times to conditions (Kratochwill and Levin, 2010); to include blinding of the patient, therapist, and assessor; to show high inter-rater reliability when recording the data (especially useful when by means of observation, Cohen, 1960); to apply the intervention as planned (see also Ledford and Gast, 2014, for a discussion on procedural fidelity); the use a repeatable measure

for the target behavior; to use an appropriate data analysis procedure; to assess generalization across other behaviors and settings; and to replicate the results.

These requirements reflect the aspects of a study or a professional practice that moderate the extent to which its findings are "solid evidence" and also affect the practitioner's confidence in the conclusions regarding intervention effectiveness. Accordingly, using a sub-optimal two-phase design such as AB (referred to as "pre-experimental," Kazdin, 1982, or "quasi-experimental," Campbell and Stanley, 1966) is a drawback, but it does not necessarily preclude a study from being useful[1], as there are other characteristics that can increase the credibility in the obtained results. In the present work, we focus on one of these aspects – data analysis – showing how to meet the condition for an appropriate data analysis.
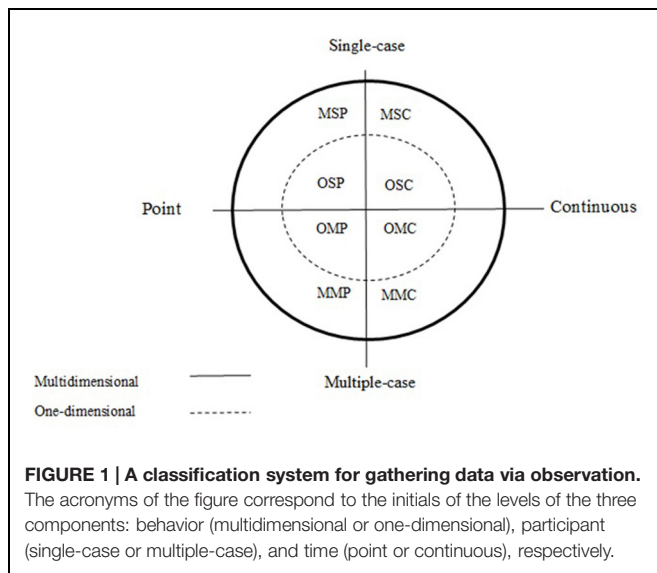
The structure of this article is as follows. First, we comment on the characteristics of non-experimental studies in order to frame a context, where improvements are required (Institute of Education Sciences, 2013). Second, we present an analytical method meeting the criterion for appropriate data analysis; we refer to its strengths, limitations, and alternatives. Third, we apply the analytical method to a real data set. Fourth, we point out several analytically challenging situations and present our own advice to practitioners and applied researchers. With the justification and illustration of the analytical method and the software, we aim to offer practitioners and applied researchers a useful tool, and indications about its alternatives.

NON-EXPERIMENTAL STUDIES

Demonstration of causal relations via experimental designs is considered optimal for building the evidence basis of interventions (Kratochwill et al., 2010; Tate et al., 2013), but everyday practice cannot always meet this requirement (e.g., due to time pressure or to the unethical withholding or removal of a potentially beneficial intervention). However, non-experimental studies can still contribute via in-depth assessment of effects, taking into consideration different sources of information (e.g., visual and numerical analyses of the data gathered, the interpretation of the client, his/her significant ones, and the practitioner) and relying on replication.

Non-experimental studies consisting only of a pre-intervention and post-intervention condition resemble "natural experiments," such as disasters or legislation changes, and they also resemble observational studies in which continuous recording of a single individual is taking place (see **Figure 1** representing the taxonomy of observation studies by Anguera et al., 2001, used in Jonsson et al., 2006). Moreover, an experimental multiple-baseline design across behaviors is similar to an observational plan in which several behaviors of the same participant are recorded each time that a video-taped situation is seen by the observers (i.e., a multidimensional observational

---

[1] Actually, even pre–post designs with a single measurement before and after an intervention can provide useful evidence (e.g., Pazzagli et al., 2014), especially if clinical significance is assessed, for instance using the Reliable Change Index (Jacobson and Truax, 1991).

FIGURE 1 | A classification system for gathering data via observation. The acronyms of the figure correspond to the initials of the levels of the three components: behavior (multidimensional or one-dimensional), participant (single-case or multiple-case), and time (point or continuous), respectively.

recording according to Anguera et al., 2001). Another similarity can be seen between a multiple-baseline design across subjects and a multiple-case one-dimensional continuous recording observational plan. However, observational (or non-experimental, in general) and experimental methodology allow reaching different conclusions. Regarding experimental control, the main differences are in: (a) the use of randomization to decide when to introduce and withdraw an intervention, (b) the staggered introduction of the intervention and (c) the replication of effects. Accordingly, in the absence of staggered introduction of the intervention, in an observational study there is less control over alternative explanations of potential behavioral change and the demonstration of intervention effectiveness is not so strong (Kazdin, 1984). Thus, multidimensional single-case continuous observation is not equivalent to multiple-baseline design across behaviors. Moreover, in a natural setting it is usually not possible to choose *at random* when to intervene in order to support internal and conclusion validity (Kratochwill and Levin, 2010). Thus, the conclusions made need to refer to the existence and amount of change in the behavior, but not to the cause for such a change.

THE ANALYTICAL METHOD EXPLAINED

The analytical method is grounded on the "data analysis" item of the RoBiNT scale: controversy remains about whether the appropriate method of analysis in single-case reports is visual or statistical. Nonetheless, two points are awarded if systematic visual analysis is used according to steps specified by Kratochwill et al. (2010, 2013), or visual analysis is aided by quasi-statistical techniques, or statistical methods are used where a rationale is provided for their suitability (Tate et al., 2013, p. 629).

Our proposal is to use the option of "visual analysis aided by quasi-statistical techniques," where the latter are understood as descriptive measures that do not intend to yield statistical significance values due to various reasons. First, visual analysis

is not only frequently used, but it is apparently the only kind of single-case data analysis that researchers seem to agree that is necessary (e.g., Parker et al., 2006; Gast and Spriggs, 2010; Kratochwill et al., 2010; Davis et al., 2013; Fisher and Lerman, 2014). Second, the evidence on visual analysis suggests that its exclusive use is potentially problematic (i.e., visual analysis is not sufficient) and techniques increasing the reliability of visual analysis are necessary (Maggin et al., 2013). Third, we consider that certain quasi-statistical techniques with favorable evidence for their performance can be used as natural complements of the commonly used visual analysis, as they share the emphasis on the same main data features (overlap, level, and trend), whereas the visual aids also take data variability into account and allow comparing projected and actual data. Fourth, applied researchers may not be willing to use the more complex statistical techniques whose results are more easily misinterpreted, in case of incomplete understanding of what exactly is being done with the data. Fifth, the use of inferential statistical procedures may not be fully justified in the absence of random sampling (Edgington and Onghena, 2007). Moreover, an inference to a population is not necessarily an aim of idiographic research (Johnston and Pennypacker, 2008) that focuses on the needs and the improvement of the individual clients. Sixth, easy to use software is available for the descriptive statistical procedures recommended here.

SYSTEMATIC VISUAL ANALYSIS

Rationale

Visual analysis has been and still is popular among professionals in their everyday psychological practice (Robey et al., 1999; Parker and Brossart, 2003) and is still advocated for (Lane and Gast, 2014) and used as a gold standard for assessing quantitative procedures (Wolery et al., 2010). Visual analysis has been considered both appropriate and sufficient for data gathered longitudinally (Michael, 1974). However, this sufficiency has been defended only for experimental studies (Sidman, 1960), which points at the need for complementing it with a quantitative procedure.

Tate et al. (2013) advise for systematic visual analysis and it necessarily starts with assessing the baseline, specifically, whether the intervention can be introduced or it should be postponed until stability is reached (Barlow et al., 2009). Alternatively, deterioration in the behavior of interest would suggest even more clearly the need for intervention. In that sense, deterioration is not expected to interfere with subsequent conclusions about intervention effectiveness (Kazdin, 1978), given that it allows exploring whether an intervention reverts the situation. Nonetheless, it is possible to assess intervention effectiveness even when the behavior is already improving before the intervention itself, as it will be shown later.

The specific data aspects, which are foci of attention, are the amount of overlap between data in the different conditions, within- and between-phase variability, slope and level change (SLC; Kratochwill et al., 2010; Lane and Gast, 2014). A more objective assessment of the degree to which data share the

same values (i.e., overlap), whether levels and trends are similar across conditions, and whether data become more stable or more variable after the intervention can be done using visual aids instead of relying on naked-eye impressions. Finally, visual analysis focuses on the whole data pattern (Parker et al., 2006) in order to assess whether it resembles the expected one, that is, a consistent improvement only during intervention. Kratochwill et al. (2010) summarize the overall assessment as a comparison between projected and actually obtained measurements. Specifically, in two-phase designs, it is relevant to project the baseline (in case it is stable or presents trend stability) into the intervention phase and compare this projection with the real treatment phase data.

Potentially Useful Tools

The assessment of overlap can be done using visual aids, such as range lines, as provided by the SCDA plug-in (Bulté and Onghena, 2012[2]) for R-Commander. The upper left panel of **Figure 2** shows an example with the data reported by Taylor and Weems (2011) for a participant called Elizabeth. This graph suggests a minor overlap between the observations. Regarding the assessment of changes in level, the same software can be used to superimpose, for instance, the median of the behavioral observations in the pre-intervention and post-intervention conditions. The upper right panel of **Figure 2** shows an example with the same data and suggests that there has been a reduction in the level of target behavior. However, the median is not very useful for the post-intervention observations in which there is a clear downward trend.

Regarding the assessment of changes in slope, two situations should be considered: when pre-intervention data are stable and when baseline data show an upward or downward trend. In case of stability, it is possible to use the stability envelope (Lane and Gast, 2014) or the two-standard deviations band used in statistical process control (Callahan and Barisa, 2005). The two-standard deviations band implies computing the average of the data for a specific condition and representing it with a solid line. The standard deviation of the same data is also computed and two dashed lines are represented: one located two standard deviations below the mean and the other two standard deviations above. The basis of this procedure is that, for a normally distributed variable, few points (less than 5%) are expected to be out of these limits in case there is no change in the behavior with the introduction of the intervention. However, we suggest using it only as visual aid and not as a formal statistical procedure, as the data cannot be reasonably assumed to be normal, continuous, or independent. This visual aid is implemented in R Core Team (2013) code[3] that only requires inputting the data and specifying the number of pre-intervention observations. As an example see the lower left panel of **Figure 2**, indicating that the reduction in behavior is beyond what is expected only by random variability as there are multiple observations with values smaller than the lower limit.

In case the pre-intervention data show a trend, it is necessary to compare the projection of this trend and the actually obtained

measurements (Kratochwill et al., 2010). For that purpose, there is another potentially useful R code[4] which allows applying the stability envelope to the trend line: (a) estimating split-middle trend (Miller, 1985), (b) projecting it into the next phase, and (c) constructing an envelope around it. The envelope can be constructed on the basis of the baseline median[5], so that the lower limit is located 25% of the median below the estimated split-middle trend and the upper limit at the same distance above it (Lane and Gast, 2014). In case 80% of the data are within those limits, this would indicate trend stability, that is, it would suggest that no change in slope has been produced with the introduction of the intervention. For using this code only data input is required before copy-pasting it in R. The lower right panel of **Figure 2** shows an example with Elizabeth's data. Given that the projected trend and its stability envelope are lower than the actual observations, this is the only piece of graphical information that does not suggest improvement in the behavior, but practitioners should be cautious when trend is estimated from as few as four observations and when it is projected farther away in time into values that are out of the range of possible measurements (Parker et al., 2011b).

Another aspect assessed is whether the introduction of the intervention has led to an immediate change in the behavior. Moreover, the duration of the change (maintained or transitory) is also taken into account in order to evaluate the strength of the intervention. A structured guide on visual analysis is offered by the What Works Clearinghouse *Standards* (Kratochwill et al., 2010; see also the application and a scoring procedure by Maggin et al., 2013) and by Lane and Gast (2014).

Limitations

Despite these guidelines on visual analysis, there are still no soundly based formal decision rules for all data aspects that are visually assessed (Kazdin, 1982) and objective and replicable outcomes are also missing (Robey et al., 1999). These two drawbacks might be among the reasons for the frequently reported inadequate performance of visual analysts (Gibson and Ottenbacher, 1988; Ottenbacher, 1990; Danov and Symons, 2008; Ximenes et al., 2009; see also Ninci et al., 2015, for a recent meta-analysis reporting insufficient interrater agreement, especially among single-case experts). Moreover, the visual analysts' decisions are not directly useful for documentation or for meta-analysis (Busse et al., 1995), which would allow establishing the evidence basis for interventions (Jenson et al., 2007), especially as generalization in single-case studies depends on replication[6] rather than on random sampling and statistical inference. As a result of these limitations, there is a consensus that visual and quantitative analyses should be used jointly (Franklin et al., 1996; Fisch, 2001; Houle, 2009; Harrington and Velicer, 2015).
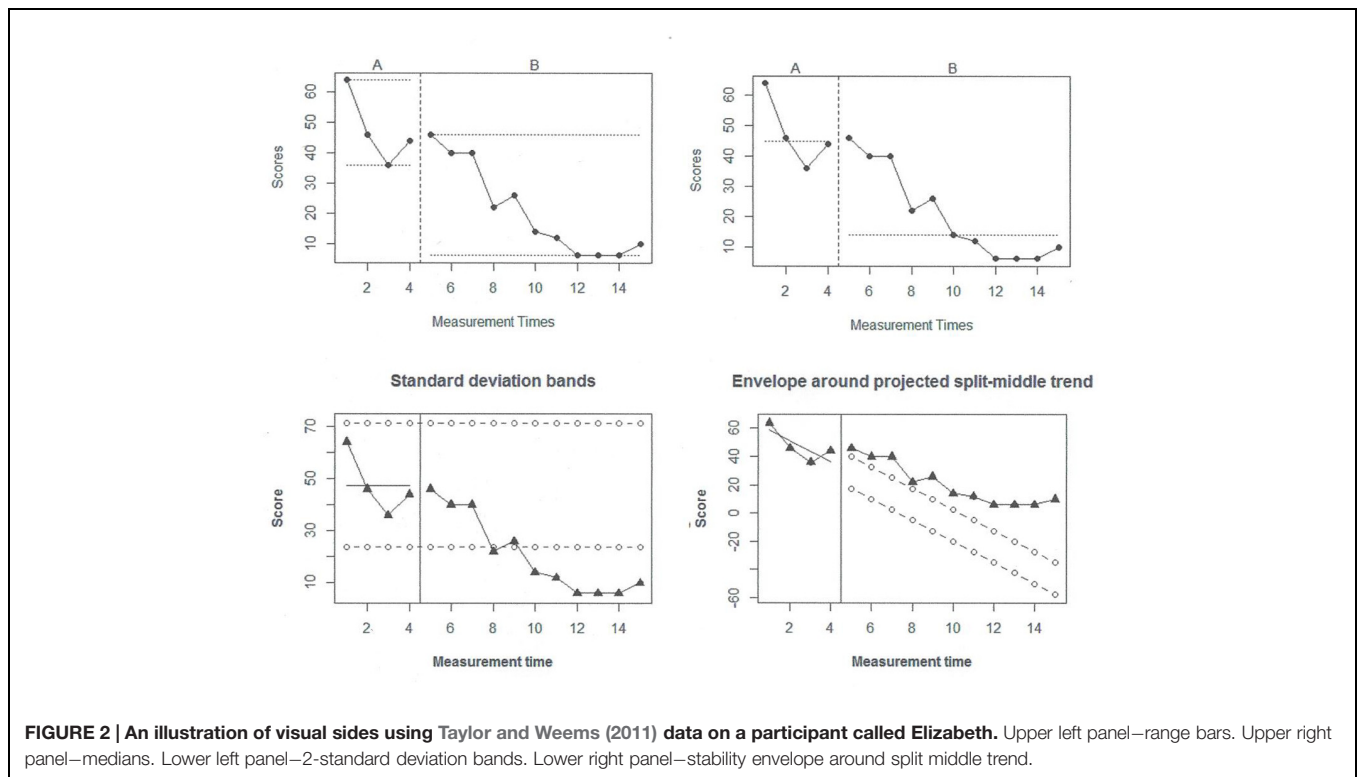
---

[2]http://cran.r-project.org/web/packages/RcmdrPlugin.SCDA/index.html
[3]https://dl.dropboxusercontent.com/s/elhy454ldf8pij6/SD_band.R

[4]https://dl.dropboxusercontent.com/s/5z9p5362bwlbj7d/ProjectTrend.R
[5]Another option is to take into account the baseline data variability, operationally defined as the interquarile range, when constructing the trend stability envelope (Manolov et al., 2014).
[6]Kratochwill et al. (2013) recommend that the findings be replicated in at least five different studies, conducted by at least three different research teams on a total of 20 participants or more (i.e., the 5-3-20 rule).

FIGURE 2 | An illustration of visual sides using Taylor and Weems (2011) data on a participant called Elizabeth. Upper left panel—range bars. Upper right panel—medians. Lower left panel—2-standard deviation bands. Lower right panel—stability envelope around split middle trend.

QUANTITATIVE ANALYSES RECOMMENDED

Our choice of procedures [non-overlap of all pairs (NAPs); Parker and Vannest, 2009 and SLC; Solanas et al., 2010a] is based on the six criteria detailed below, although alternative quantifications are provided later in this article.

Criterion 1: Simple to Compute

The techniques are relatively simple to compute and offer straightforward interpretations for practitioners who are not experts in statistics (as the Institute of Education Sciences, 2013, suggests). The calculation does not entail statistical decisions about the likelihood of obtaining such a large difference under the null hypothesis. This criterion also relates to the need for easily trainable procedures (Fisher et al., 2003).

Criterion 2: Complementary to Visual Analysis

This criterion is related to the popularity of visual analysis among practitioners (Parker and Brossart, 2003), which makes necessary to develop and promote suitable complements to it. NAP and SLC are actually based on relevant visual criteria (i.e., data overlap, change in slope and in level) and thus potentially useful as complements[7]. Specifically, visual inspection can be used to assess

the adequacy of the baseline as a reference for comparison. The change identified visually can then be quantified in an objective manner. The numerical values also offer information that can be communicated among researchers and professionals and used for further analyses with different analytical techniques or as part of research synthesis (e.g., NAP was used in the meta-analysis by Jamieson et al., 2014, whereas the new developments on SLC make possible its comparability across studies; Manolov and Rochat, 2015).

Criterion 3: Synergic Application

Wolery et al. (2010) criticized non-overlap methods for omitting relevant data aspects such as level, trend, and stability or variability: SLC partially addresses this issue and it also responds to Beretvas and Chung's (2008) suggestion for quantifying separately level and slope change. Moreover, SLC yields unstandardized results, which help assessing the practical importance of the behavior change when using meaningful measures (Grissom and Kim, 2012) such as the number of tantrums or the number of self-injurious behaviors. In contrast, NAP is bounded, which allows comparisons and quantitative integrations. Thus, NAP and SLC can be used jointly as they provide different information. Specifically, NAP is an ordinal measure (Solomon et al., 2015) that does not distinguish between conditions once complete overlap is achieved. In contrast, SLC can be used even in absence of overlap to quantify how different the measurements belonging to different phases are.

---

[7]Wolery et al. (2010) found that no overlap technique had highest agreement with visual analysts for both data with and without a change. However, they did not include NAP or Tau-U (Brossart et al., 2014) in their study, and these two non-overlap indices are considered to be superior, given their more solid statistical basis

and greater statistical power according to the review performed by Parker et al. (2011a).

Criterion 4: Absence of Assumptions and Restrictions of Use

The procedures used here do not make explicit *a priori* assumptions about independence or homoscedasticity of the data, as serial dependence is likely to present in data obtained from the same individual (Matyas and Greenwood, 1996). There are also no specific design requirements.

Criterion 5: Appropriate Performance

In relation to the previous point, there is evidence that their performance is appropriate for a variety of single-case data patterns (Manolov et al., 2011). NAP is a suitable indicator when data is stable and even when data is variable. In contrast, in such situations visual analysis is more difficult to perform and means and medians are not informative and trends are not estimated with precision. On the other hand, NAP is not suitable when the data show improving trend, but SLC can be applied in such a situation – this complementarity relates to Criterion 3 "Synergic application." SLC is useful for separately quantifying the change in level and the change in slope in potentially meaningful terms. In relation to this criterion, it is important to discourage the use of methods for comparing conditions that have been shown not to perform appropriately, such as the binomial test applied after the split-middle method (Crosbie, 1987) which does not control for Type I error rates, ITSACORR which presents modeling flaws (Huitema et al., 2007), or the C-statistic (Young, 1941; Tryon, 1982; used by Fabio et al., 2013), which is actually an estimator of autocorrelation (DeCarlo and Tryon, 1993).

Criterion 6: Reduced Likelihood of Misinterpretation

Using descriptive measures like the ones provided by NAP and SLC makes it less likely for applied researchers to make inferences, which would be statistically incorrect in absence of random sampling of the participant or of the behavior of interest (Barlow et al., 2009). We consider that inferential statistical techniques are more susceptible to being misunderstood and to prompt researchers to make dichotomous decisions (Cohen, 1994) about intervention effectiveness or behavioral change. In case inference is desired, we recommend causal inference, instead of population inference, in line with the recommendations by Heyvaert et al. (2015).

Non-overlap of All Pair

Non-overlap of all pairs is an improvement of the Percent of non-overlapping data commonly used for quantifying the degree to which the measurements pertaining to each phase share the same values (Scruggs and Mastropieri, 2013). It represents the number of non-overlapping data relative to all possible comparisons and it is actually identical to the non-parametric version of the probability of superiority (Grissom, 1994), which is related to the common language effect size (McGraw and Wong, 1992). When a decrease in the behavior is expected, as in the example provided later, the formula for this indicator can be written

as $(\#(X_{pre(i)} > X_{post(j)}) + 0.5\#(X_{pre(i)} = X_{post(j)}))/n_{pre}n_{post}$ where X_{pre} and X_{post}, which represent the values of the pre-intervention and post-intervention phases, respectively, with $i = 1, 2, \cdots, n_{pre}$ and $j = 1, 2, \cdots, n_{post}$, and # denotes the number of times that the inequality or the equality is true. Given that each data point of the pre-intervention phase is compared to a data point from the post-intervention phase there is a total of $n_{pre}n_{post}$ comparisons, where n_{pre} and n_{post} denote the number of measurements in the first and second phase, respectively. In each of these comparisons, a non-overlap occurs when a post-intervention measurement represents an improvement over a pre-intervention measurement, with ties counting as half a non-overlap. To obtain the index value, the number of non-overlapping pairs is divided by number of comparisons. This value can be interpreted in two different ways. One the one hand, it represents the proportion of comparisons for which intervention phase data improve baseline data. On the other hand, it can be conceptualized as the probability that a randomly selected post-intervention data point will improve (here, be smaller than) a randomly selected pre-intervention data point. The NAP can be computed via the online calculator http://www.singlecaseresearch.org/calculators/nap by Vannest et al. (2011), where it is only necessary to enter the data from the different conditions in separate columns. It is also part of the output ("A vs. B" comparison) of the R code for Tau-U https://dl.dropboxusercontent.com/u/2842869/Tau_U.R (Brossart et al., 2014), which requires loading a data file with a single comma-separated column including "Time" (1, 2, ..., $n_{pre}+n_{post}$), "Score" (denoting the measurements) and "Phase" denoting the condition (n_{pre} times the value of 0 followed by n_{post} times the value of 1).

Slope and Level Change

Slope and level change quantifies two aspects of behavior's evolution after a change in the conditions: change in slope and change in level. Actually, this procedure first estimates pre-intervention linear trend ($\widehat{\beta_A}$) as the average of the differenced first phase measurements, that is, $\widehat{\beta_A} = \sum_{i=1}^{n_{pre}-1}(X_{i+1} - X_i)/(n_{pre} - 1)$. Baseline trend is thus the average increase (or, if negative, decrease) from one baseline measurement occasion to the next one. This estimation can inform about the characteristics of the data before an intervention is introduced. Moreover, baseline trend is removed from the whole data series so that it does not affect the quantification of the effects of the intervention. Technically, each data point is corrected according to its position in the series of observational sessions. This initial step allows for applying an intervention even when the theoretically undesirable linear improvement is present already during the assessment period. Thus, SLC would show whether there is an effect of the intervention beyond the initial improvement. After the correction it is assumed that the pre-intervention phase shows zero trend (i.e., stable data) and thus the trend present in the post-intervention phase actually represents an effect (i.e., a change in slope). This effect is estimated in the same manner as in the initial step, that is, as the average of the differenced

(and already detrended) post-intervention measurements: $\widehat{SC} = \sum_{j=1}^{n_{post}-1} (\widetilde{X}_{j+1} - \widetilde{X}_j)/(n_{post} - 1)$, where $\widetilde{X}$ represent detrended values (i.e., after eliminating pre-intervention trend), instead of the original measurements. Therefore, the intervention phase estimate of trend presents the average increase (or, if negative, decrease) from one intervention phase measurement occasion to the next one, after controlling for baseline linear trend. For instance, the slope change estimate reflects the average decrease in the number of tantrums in a child with each successive post-intervention measurement, that is, a progressive change.

Once slope change is estimated, post-intervention trend is removed in order to obtain a net estimate of the change in level. This way of proceeding is similar to what is done in ARIMA models, before obtaining a quantification of change in level (see Harrington and Velicer, 2015). Net change in level is estimated as the difference between the average of the corrected post-intervention measurements and the average of the corrected pre-intervention measurements. The expression for this step is $\widehat{LC} = \sum_{j=1}^{n_{post}} \widetilde{X}_j/n_{post} - \sum_{i=1}^{n_{pre}} \widetilde{X}_i/n_{pre}$, where $\widetilde{X}$ represents post-intervention measurements with both pre-intervention trend and post-intervention trend (i.e., slope change) removed and $\widetilde{X}$ represents pre-intervention measurements with pre-intervention trend removed. The net level change estimate quantifies, for instance, the average decrease of tantrums in a child after the intervention, once slope change has been taken into account. Thus, it can be conceptualized as a quantification of an abrupt and maintained effect. The SLC can be computed using R code https://dl.dropboxusercontent.com/s/ltlyowy2ds5h3oi/SLC.Ror via the R-Commander Plug-in offering point-and-click menus, available at http://cran.r-project.org/web/packages/RcmdrPlugin.SLC/index.html. For obtaining the numerical results and a graphical representation of the original and detrended data, both options only require inputting the values of the observations and specifying the pre-intervention phase length.

ALTERNATIVES FOR QUANTITATIVE ANALYSIS

There is currently no consensus on which the optimal quantitative procedure for single-case designs is (Kratochwill et al., 2010; Smith, 2012), as the RoBiNT scale also reflects (Tate et al., 2013). For a comprehensive review of most currently available techniques the interested reader should consult the state-of-the-art information provided in the Special Issues of the *Journal of School Psychology* in 2014, volume 52, issue 2 (e.g., Shadish et al., 2014; Swaminathan et al., 2014) and of *Neuropsychological Rehabilitation* also in 2014, volume 24, issues 3-4 (e.g., Borckardt and Nash, 2014; Brossart et al., 2014; Heyvaert and Onghena, 2014). Here, we provide brief comments on the strengths and limitations of several analytical alternatives, which in some cases may be more appropriate than NAP and SLC included in the analytical method suggested.

Considering specifically observational studies in which data is recorded continuously within a session, it is possible to follow an analytical approach different from the one used in single-case designs, namely, to apply sequential analysis to explore whether the occurrence of some behaviors make more or less probable that other behaviors take place (Bakeman and Quera, 2011). Additionally, longer series of data gathered across time can be analyzed using Markov chains or analyses of rhythm, according to the aims of the study (Suen and Ary, 1989).

Starting our discussion from procedures similar to the ones included in the analytical method, Tau-U (Parker et al., 2011b) is closely related to NAP and it is preferable when pre-intervention trend is present in the data. For both Tau-U and NAP *p*-values have been offered, although their basis has not clearly been explained in the presence of autocorrelation. However, Tau-U is interpretatively and computationally less straightforward than NAP (i.e., Criterion 2 "Complementary to visual analysis" is met to a lesser extent). For instance, even in case a baseline trend is generally deteriorating, if there is a single improving value in the baseline phase, as compared to a previous baseline data point, this would reduce the value of the non-overlap index. Thus, in case trend is not reasonably clear, Tau-U can be an excessively conservative procedure (i.e., it would overcorrect). Furthermore, more evidence is required on its performance (thus the abovementioned Criterion 5 "Appropriate performance" is not fully met, as Parker et al., 2011a,b, offer only applications to real data, but no simulation study).

Regarding procedures quantifying average differences, similar to the SLC, the *d*-statistic (Shadish et al., 2014) has to be mentioned. We highlight here the *d*-statistic developed by Shadish et al. (2014), which has been created specifically for single-case designs rather than the *d*-statistic described by Busk and Serlin (1992; approach one[8]), recommended by Beeson and Robey (2006), for two reasons: (a) the latter is an adaptation of the group designs indicator and does not take into account autocorrelation, while it has been shown to be somewhat affected by autocorrelation (Manolov and Solanas, 2008); and (b) its sampling distribution in single-case studies is unknown (Beretvas and Chung, 2008). In contrast, the *d*-statistic developed by Shadish et al. (2014), offers a standardized measure of the mean difference with a solid statistical basis offering the possibility to estimate the index variance for future meta-analyses. So far, it has been developed for AB, reversal (e.g., ABAB) and multiple-baseline designs and assuming that pre-intervention data is stable, assuming that within-case residuals and between-case variation do not change over time. Thus, this procedure fails in terms of Criterion 4 "Absence of assumptions and restrictions of use." Some potential drawbacks include: (a) its computation requires several cases per study; and (b) the calculations are potentially difficult to understand by applied researchers with less statistical knowledge and require the use of software, such as the R code provided in the appendix of the Shadish et al. (2014) paper. Hence, the *d*-statistic is preferable to SLC when there is more than one participant per study and the aim is to obtain a standardized

---

[8]This indicator is equivalent to Glass' Δ (Glass et al., 1981), as it divides the mean difference by the standard deviation of the pre-intervention phase data.

measure, but it is not suitable when pre-intervention trend is present and when the focus on a specific client.

Generalized least squares regression analysis (Swaminathan et al., 2014) also enables computing an effect size index. Its strengths include the fact that it can take into account changes in level and in slope (although they are quantified as part of the same overall indicator, unlike SLC), the versatility in modeling (e.g., controlling for linear and non-linear trends), and that it deals explicitly with autocorrelation. However, autocorrelation estimation has been shown to be problematic (Solanas et al., 2010b) and the analytical procedure requires several steps, some of them taking place iteratively (i.e., Criterion 1 "Simple to compute" is not met). This procedure is applicable to longer data series for which autocorrelation can be estimated with greater precision. Moreover, we recommend that practitioners work together with a statistician, so that the analysis can be properly run. Brossart et al. (2006) compared the agreement between visual analysis and several regression-based approaches and the best performer in this terms (related to Criterion 2 "Complementary to visual analysis") was Allison and Gorman's (1993) method, which is however affected by autocorrelation (Manolov and Solanas, 2008). The generalized least squares approach was not yet proposed by the time Brossart et al. (2006) conducted their study and more evidence is necessary to assess its performance.

Multilevel models are an extension of piecewise regression and can be used to model several data aspects (e.g., trend, autocorrelation, heterogeneous data variability across phases) and they yield estimates of the change in the same measurement units as the target behavior and their statistical significance (Moeyaert et al., 2014a). The main drawbacks of multilevel models are the problematic estimation of variance (Ferron et al., 2009), their relative complexity for applied researchers with less statistical knowledge and the fact that they the replication of the intervention in several participants. Actually, such a complex procedure is more suitable for more complex design structures that the two-phase AB (Moeyaert et al., 2014b). Finally, most implementations of this analytical procedure have been done in commercial software (e.g., Moeyaert et al., 2014a include SAS code in their article).

An effect size index can also be computed from interrupted time series analysis via ARIMA (autoregressive integrate moving average) models, which allow controlling for trend and autocorrelation (Simonton, 1977). The main difficulties of this option are the need for long data series and the problematic initial model identification step. However, there have been suggestions for using some general models that make model identification unnecessary (Harrop and Velicer, 1985). A recent application of ARIMA models has shown that these can be applied to two-phase data, but there might be convergence problems and, more importantly, the agreement with visual analysis is low (Harrington and Velicer, 2015). We consider that this latter drawback and the relative complexity of the technique make it less attractive to applied researchers with no statistical expertise.

Statistical significance (i.e., p-values) can be estimated for d and the generalized least squares procedure on the basis of the comparison between the test statistic and a theoretical reference (the sampling distribution) and allows making inference about the population from which the individual was drawn. In contrast, randomization tests (Heyvaert and Onghena, 2014) yield a p-value on the basis of a comparison between the test statistic and an empirical reference –the randomization distribution. In the current context of two-phase studies, this reference is the distribution of the test statistic values quantifying the difference between the two conditions for each possible intervention start point (i.e., for each possible way in which the data series can be split into two; Edgington, 1980). For this analytical option the inference is restricted to the case studied, referring to the likelihood of obtaining such a large difference in case the intervention was ineffective. Randomization tests are versatile in terms of test statistic to use (e.g., it can be an effect size such as a non-overlap index) and offer flexible options for dealing with different situations (e.g., Levin et al., 2012). However, the necessary randomization as part of the data collection process is both a strength (Kratochwill and Levin, 2010) and a limiting characteristic (Fisher and Lerman, 2014) in a clinical setting (i.e., Criterion 4 "Absence of assumptions and restrictions of use" is not met). Moreover, in certain conditions Type I error rates are not controlled (Manolov et al., 2010). Randomization tests can be recommended when the aim is to obtain statistical significance and the point(s) of change in the conditions can be chosen at random. Randomization tests are also accompanied by freely available software (Bulté and Onghena, 2013; Levin et al., 2014).

Another procedure using an empirical reference distribution is simulation modeling analysis (SMA; Borckardt and Nash, 2014). In SMA, data are generated with the same autocorrelation as estimated from the data, but with no difference between the conditions, thus representing the null hypothesis of identical behavioral level across conditions. The p-value represents the likelihood of the outcome, computed as a point biserial correlation between the measurements and a dummy variable representing the condition (0 = without intervention, 1 = with intervention). This approach is intuitive, takes autocorrelation into account, and it can be implemented via the software available freely at http://clinicalresearcher.org/software.htm. However, so far the evidence on its performance (i.e., Criterion 5 "Appropriate performance") is not sufficient. Finally, as the focus of is put on the p-value, which may enter in conflict with Criterion 6 "Reduced likelihood of misinterpretation."

Whereas SMA uses Monte Carlo methods or bootstrap for generating samples and estimating the likelihood of the value of test statistic in case there is not difference between conditions, bootstrap has also been suggested for single-case as a way of reducing bias and estimating standard errors (McKnight et al., 2000) and specifically for estimating confidence intervals of regression-based R-squared values (Parker, 2006). This option has not received much attention lately and it is unclear whether applied researchers would be willing to use it.

Another computer-intensive option could be the Monte Carlo based method for modeling non-linearity proposed by Theiler et al. (1992). However, modeling non-linear patterns can also be achieved without prior knowledge and without the need to specify a model, by using local regression (LOESS; Jacoby,

2000; Solmi et al., 2014). We consider LOESS to be more practical for applied researchers than the Theiler et al. proposal. Moreover, randomization tests are also more parsimonious as they require no assumptions about the process generating the data or about random sampling. Actually, Theiler et al. (1992) mention this option as rank statistic approach for obtaining p-values. Randomization test offer the advantage of not only mimicking the preserved data features (such as mean and standard deviation), as expressed by Theiler et al. (1992), but they actually preserve the whole data series and its order, taking advantage of the different possible moments of change in phase, when such moments are determined at random.

A simplified summary of these general recommendations regarding the use of the analytical techniques can be found in **Figure 3**.

INTERVENTION EFFECTIVENESS IS NOT ONLY DATA ANALYSIS

Assessing the relevance of an intervention cannot be constrained solely to visual and descriptive or inferential statistical analyses. It is important to assess aspects such as quality of life (Kendall, 1999), whether the behavior has moved from dysfunctional to functional ranges (Kazdin, 1999), without forgetting subjective evaluation (Hugdahl and Ost, 1981). Regarding the latter, Kratochwill and Levin (2010) highlight the need to get to know the perceptions of the client and of significant others. According to the specific context being studied, these significant others would be the family members (parents, siblings, marital partner), the teacher, the coach, or the boss (as figure with a higher hierarchical role), and friends, classmates, or colleagues (at the level of "peers"). Kazdin (1984) has referred to these groups of people as "paraprofessionals," as they help detecting the behavior that requires intervention and they can also be the agents reinforcing the behavior of interest (e.g., a mother reinforcing a child's disruptive behavior by paying attention to it) or producing stimuli for discriminating conditions in which certain types of behavior are desirable (e.g., a boss may encourage jokes with one type of clients and more distant behavior with others).

THE ANALYTICAL METHOD APPLIED

In the present section, we will illustrate the application of the analytical method and the information that can be obtained via visual and quantitative analyses, while also considering substantive criteria. This application focuses on the family context, where it is common to gather data before and after an intervention (Crane, 1985). One of the empirically supported interventions in this context is the Parent Child Interaction Therapy (PCIT; Eyberg et al., 2008), which has been reported to increase positive parent behavior and reduce child behavior problems (Borrego et al., 2006). For the current example, the data gathered by Bagner et al. (2009) will be used. The participants are a 23-months-old premature-born child displaying difficult behaviors and his mother. The application of the PCIT focuses on
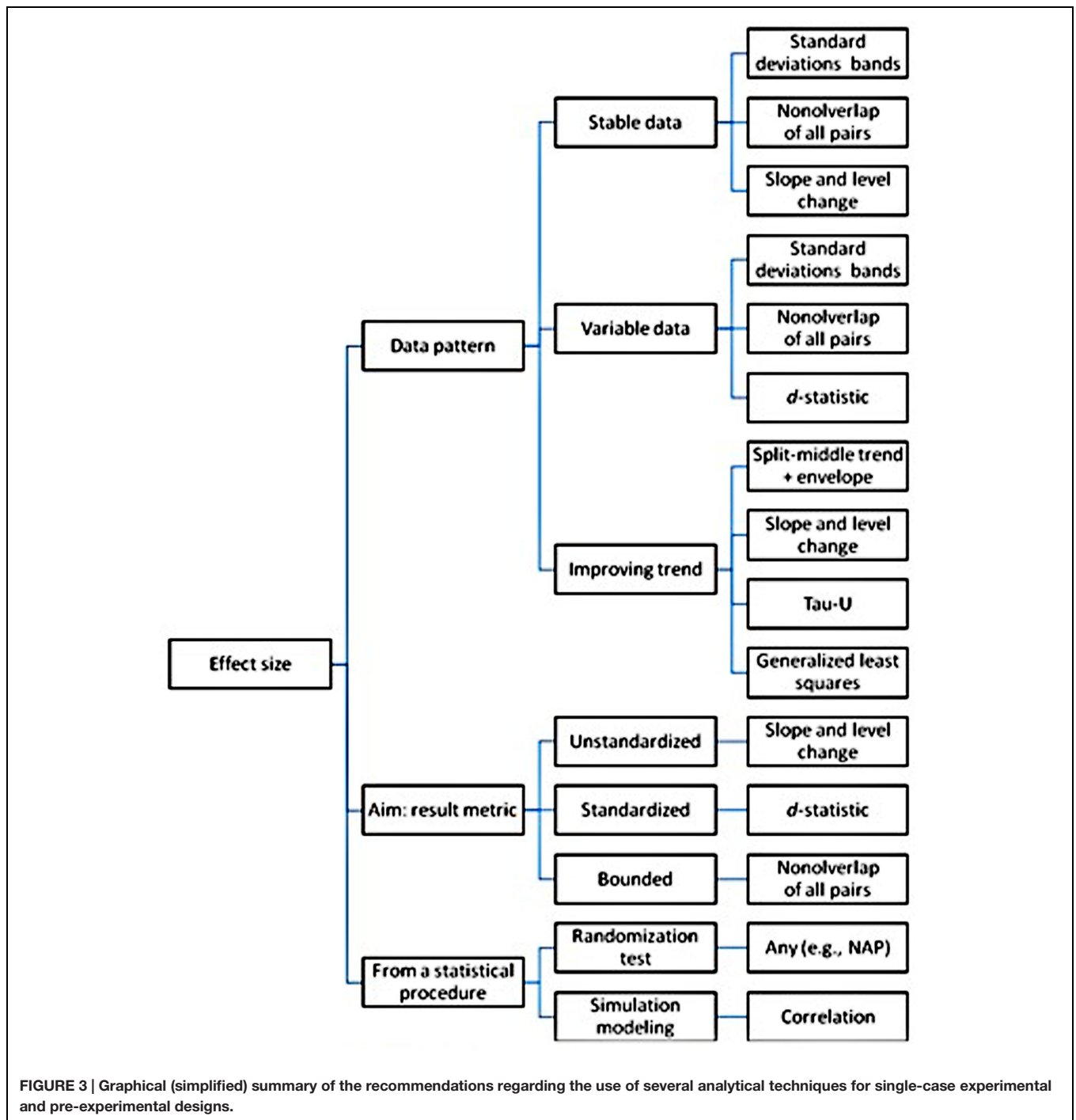
teaching parenting skills in order to improve the interaction with the child and to decrease his externalizing behavior. Teaching takes place in two phases. First, child-directed intervention (CDI) takes place. It is similar to play therapy: the child is the leader and the parent has to learn how to act positively (e.g., praising the child, imitating the child's play). Second, parent-directed intervention (PDI) phase occurs. It is similar to clinical behavior therapy: the parent is more directive and has to improve her way of disciplining so that a greater compliance is achieved. In order to assess intervention effectiveness, several sources of information are used: parent reports provided via inventories, observation of the parent–child interaction, and physiological measurements. In the running example, we focus on the parent weekly reports obtained via the Intensity scale of the Eyberg Child Behavior Inventory (ECBI; Eyberg and Pincus, 1999) on disruptive behavior, although a complete assessment entails exploring whether all available information converges to the same conclusion. The Bagner et al. (2009) ECBI data were chosen here given that there is a cut-off point at a T-score of 60 which indicates clinically significant results and eases the interpretation in substantive terms. The data gathered[9] on the ECBI scale are represented on **Figure 4**. The upper panel contains ordinary least squares trend lines provided by the SCDA plug-in for R, the middle panel contains split-middle trend for the first phase, and the lower panel represents the application of the two-standard deviations band fit to the first condition's data and projected into the second one.

Firstly, when visually inspecting the data, it has to be kept in mind that both phases are treatment phases and thus in both some reduction in child's behavior is expected and desired. Moreover, it has to be taken into account that the pre-treatment (i.e., actual baseline) value is 82, equal to the first CDI phase measurement. At the beginning of the first phase there is actually a reduction, but then a new increment starts. Considering this alternating pattern the CDI does not seem especially effective. Given the amount of variability in the first phase, neither the central tendency measure (mean represented on the lower panel of **Figure 4**), nor the different types of trend fitted (upper and middle panel) seem to represent the data well-enough. This can hamper the comparison between this condition and the subsequent one.

Once the intervention is introduced, there is apparently a decrease in the ECBI score on disruptive behavior. The downward trend is stable, as shown by the good fit of the ordinary least squares regression line to the data (upper panel of **Figure 4**). For such data it is not meaningful to discuss level or variability around a mean or a median level; actually variability is only assessed looking at the (small) distance of the measurements from the fitted trend line.

Comparing the two phases in terms of overlap, the values in the beginning of the PDI-phase are similar to the ones in the CDI-phase, but not so in the end. Comparing levels is not meaningful. Comparing trends is hindered by the lack of fit of the trend lines to the CDI data, but if we focus on the last four (out of five)

---

[9]We would like to thank Dr. Daniel Bagner for kindly offering the raw data for re-constructing their original figure.

FIGURE 3 | Graphical (simplified) summary of the recommendations regarding the use of several analytical techniques for single-case experimental and pre-experimental designs.

CDI measurements, there is a deterioration that is reverted with the introduction of the PDI: thus a change in slope has taken place. The comparison between projected and actual data is done in two ways, projecting the baseline mean with limits based on the baseline standard deviation and projecting the split-middle trend line with limits based on 25% of the baseline median. In this case, both approaches lead to a very similar graphical representation, which is well-aligned with the conclusion that the last PDI data points are clearly lower that what would
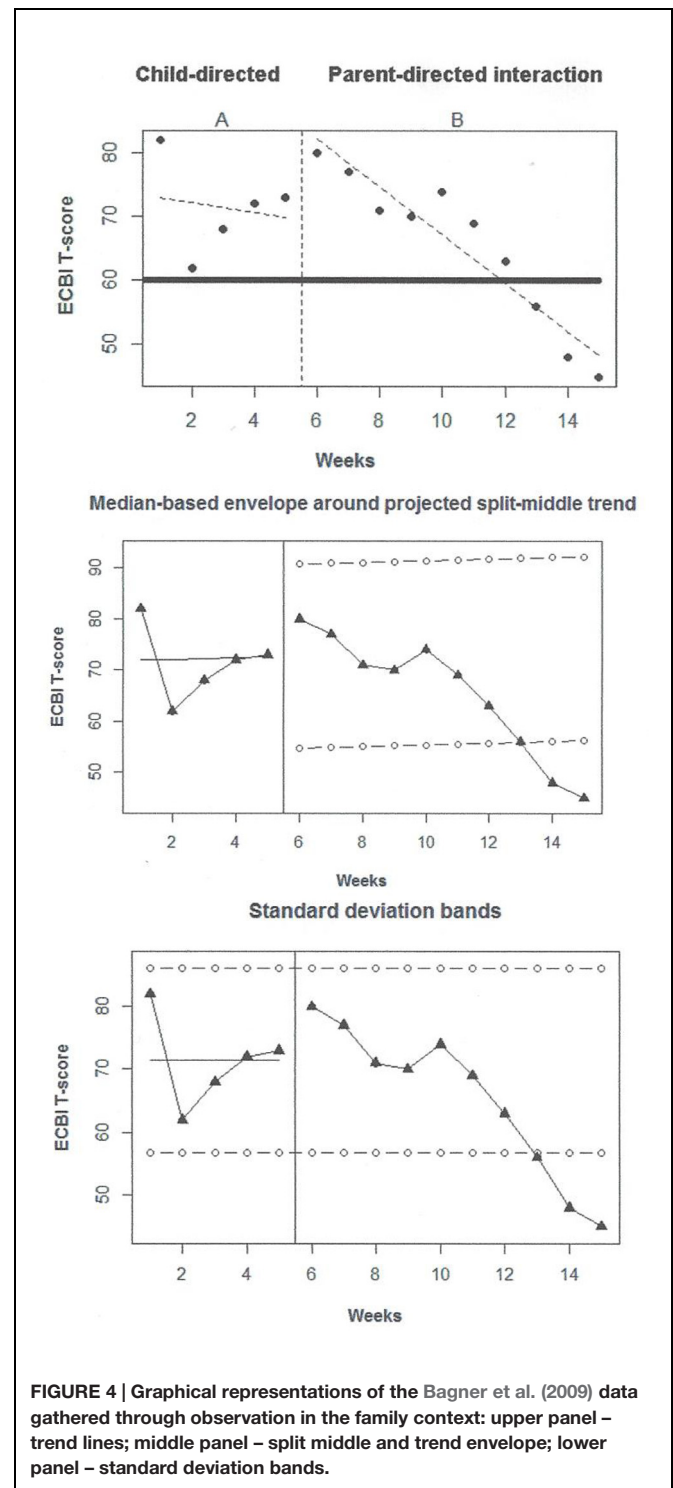
be expected (i.e., values within the limits) in case there was no difference between the two interventions. Additionally, we should consider that Bagner et al. (2009) collected a post-treatment measurement equal to 38 – a value even lower than the last PDI-phase measurement and so the downward trend seems to continue, which could be interpreted as maintenance of the effect.

Secondly, regarding quantitative analyses, the NAP performs 50 comparisons, given that $n_{pre} = 5$ and $n_{post} = 10$, in which

there are 19 full overlaps, that is, 19 cases in which a CDI datum is better (here, lower) than a PDI measurement, 0 ties, and 31 cases in which a PDI measurement is better than a CDI data point. (Lower rather than greater values are considered as overlaps, given that the aim is to reduce the disruptive behavior and thus also the ECBI T-score.) The value yielded by NAP is 62.00%, which can be interpreted as the percentage of PDI measurements that improve the CDI measurements. Therefore, the index does not suggest that the change is especially salient, given that the value is only slightly higher than the one expected by chance (50%) and it is within the range of values (0–65%) denoting small effect according to Parker and Vannest (2009). However, it has to be considered that this may be due to the fact that the effect is delayed. The data pattern is not specifically easily analyzed by the SLC either. The procedure estimates the CDI-phase trend as −2.25, which represents an average of approximately two T-score units reduction for each CDI measurement time. However, this value does not reflect the visual impression, provided that this phase shows a specific kind of variability (i.e., an alternating pattern). Correcting for this initial phase trend, the slope change estimate is −1.64, that is, nearly two T-score points average reduction for each PDI measurement time. This quantification reflects to some extent the visual impression of slope change. SLC's estimate of the net change in level is positive, 18.15, which contrasts with the visual impression of the graphed data.

Thirdly, focusing on substantive criteria, Bagner et al. (2009) summarize their results in terms of improved parent practice and increased child compliance. In fact, while the former result stems from observation and evaluation by the authors, the latter is based in reports from the parents (i.e., the paraprofessionals). Regarding the ECBI scores, the last three scores during the PDI phase fall out of the clinical range, indicating that a practically significant change in behavior of the child has taken place. Interestingly, these same three scores also fall out of the two-standard deviations band and out of the split-middle trend stability envelope represented in the middle and lower panels of **Figure 4**. To complement this assessment, the authors report that at a 4-months follow-up the results of the ECBI remained in the normal range (the value was 47), which increases the confidence in the importance of the behavioral change. Finally, it should be noted that Bagner et al. (2009) comment explicitly the "inability to conduct statistical analyses" (p. 475), which suggests that informing applied researchers about analytical options for two-phase single-case designs, as we intend with the current paper, is a timely endeavor.

The main conclusion of this application of the analytical method is that visual analysis is necessary for focusing at different aspects of the data, such as an unstable baseline which is not well-represent by mean or trend lines, a somewhat delayed slope change, and a considerable amount of overlap only in the beginning of the second condition but not at the end. The variability and relative shortness of the first phase (although it meets the current standards of five measurements; Kratochwill et al., 2010) have to be kept in mind when comparing it to the measurements obtained in the subsequent condition. In the current case, the visual aids reflected this

FIGURE 4 | Graphical representations of the Bagner et al. (2009) **data gathered through observation in the family context: upper panel – trend lines; middle panel – split middle and trend envelope; lower panel – standard deviation bands.**

variability and suggested a similar conclusion as the one based on substantive criterion expressed as a cut-off point. All this information is critical for interpreting correctly the numerical yielded by descriptive statistical procedures. Actually, we preferred to use a data set that is challenging for the quantitative analyses in order to alert applied researchers

on the need to interpret numerical values with caution and to use all information available; we also wanted to avoid doubts about the data being picked up only to show the quantification in a positive way (Fisher and Lerman, 2014). Finally, the follow-up measures, the parent-report and the physiological measures recorded by Bagner et al. (2009) also contribute to building solid conclusions. The two-phase design may not be sufficient for establishing a causal effect in a scientifically sound way, but there is enough information pointing at the clinically important reduction of problematic behavior.

DISCUSSION

The present work focused on the question of what can be done to improve the data analysis in studies/practices using sub-optimal designs in such a way that results are more useful to the discipline. We recommended an analytical method consisting of structured visual analysis complemented with descriptive statistical procedures, while also keeping in mind substantive criteria (i.e., the opinion of the individuals involved in the process: family members, teachers, peers, coworkers, or supervisors). On the one hand, quantifications are useful for summarizing different aspects of the data and making the results available for subsequent meta-analysis. On the other hand, visual analysis is required for gaining an in-depth knowledge of the data and for assessing the adequacy of any specific quantitative procedures, due to the lack of consensus regarding the most appropriate technique (Tate et al., 2013).

A second question concerned the availability of tools for implementing the procedures proposed as part of the analytical method. We have mentioned, referenced, and illustrated the output of several tools implemented in the freeware R. Some of them are based on clickable menus, whereas others only require inputting the data before copying and pasting the code. The availability of software is crucial for eliminating the errors in obtaining the numerical and graphical results and in terms of time efficiency, both for short and relatively straightforward data series (e.g., Bunn et al., 2005) and for longer series with and less visually clear data patterns (e.g., Abney et al., 2014).

One potential issue with the analytical method is that it is possible that, in some instances, the three components do not coincide. A cautious approach would be to gather follow-up data after a certain period of time in order to check whether the initial ambiguous result of the assessment still holds. In case the unclear change is maintained and perceived as a change by the participants, then there would be evidence in favor of its practical importance. If there is disagreement between the substantive criterion and the other two components, we think that if the clients' well-being, quality of life, functionality, performance, etc. is improved according to their own opinion, then the substantive criterion should prevail, regardless of its numerical expression. In any case, the general effectiveness of an intervention depends on replications (Pashler and Wagenmakers, 2012) and not on the numerical result in a single study. Finally, if there is a divergence between the visual and quantitative information, it is

important to know: (a) whether there is any data feature (e.g., pre-intervention trend, outliers) that might affect the performance of the quantitative analysis – in such case visual inspection should prevail; or (b) whether the data pattern prevents from getting a clear visual impression (e.g., due to highly variable data and/or a complex design structure) – in such case the quantitative summary is potentially more useful.

Another issue with the analytical method is that it might fail in certain situations such as the ones described in this paragraph (the list is not necessarily comprehensive). First, it is possible that the pre-intervention phase is too short or the measurements too variable for estimating trend with precision: the SLC quantifications would be less useful, but if there is no clear evidence of trend, then the NAP can be used as main quantification. Second, if there is complete non-overlap between the observations of the two conditions, the NAP will not be very informative, but the SLC can be used as an unstandardized quantification of the amount of difference and the d-statistic as a standardized quantification if more than one participant is being studied. Third, there might be a non-linear trend present in data, which is not an optimal situation for applying the SLC. In such case running medians (Tukey, 1977) can be used as a visual aid via the SCDA plug-in for R, while data modeling via the generalized least squares approach and LOESS is also possible. Fourth, there might be a delayed change in the behavior, not occurring simultaneously with the change in conditions (an issue that has remained practically unstudied except for Lieberman et al., 2010). In such case, the descriptive statistics will reflect the delay with lower quantifications of the effect, but it would be crucial to explore the cause of the change among the external uncontrolled factors (i.e., the solution is not an analytical one), given that the immediacy of the effect is one of the cornerstones for demonstrating causality (Kratochwill et al., 2010).

We hope that the discussion presented here would help practitioners and applied researchers to apply a systematic approach to data analysis and take a step toward partially improving the methodological quality of the studies. However, this would only be *one* step and studies would also need to meet the recommendations about the assessment and measurement of the target behavior, the implementation of the intervention, and the use of blinding to ensure objectivity, and also about reporting the results of the study (Tate et al., in press). Finally, it should always be considered whether what is assessed can be considered an "intervention effect" (in causal terms) or only a "behavioral change," which after several replications might point at the possible effectiveness of the intervention. In that sense, the analytical method was described in the context of studies with less-than-optimal designs in which causal relations cannot be readily established. Nonetheless, it is possible to extrapolate the method to experimental situations (e.g., multiple-baseline designs in which it is crucial to assess whether the behavioral change coincides with the staggered introduction of the intervention).

As a limitation of the quasi-statistical component of the analytical method, it is debatable whether the numerical results can be presented confidently in absence of a conventionally

accepted optimal procedure, i.e., when all analytical techniques can be criticized. Considering the analytical method as a whole, further discussion is necessary on how to proceed when practitioners are faced with data that cannot be easily analyzed visually or quantitatively (e.g., short series, great data variability). One option would be to use the substantive criteria as basis for the conclusions and label the study as "practice" but not as "research." In contrast, when all three pieces of information (visual, quantitative, and substantive) coincide, it still has to be kept in mind that not meeting current *Standards* (Kratochwill et al., 2010) could render two-phase studies only a "pilot" status and, when included in meta-analysis, they are likely to be assigned lower weights and have less influence on the summary measures obtained.

AUTHOR CONTRIBUTIONS

The initial idea was due to JL and it was subsequently complemented and further developed by RM. The manuscript was written by JL (observational, non-experimental conceptual part in the Introduction) and RM (analytical part in the Analytical Method Explained, Analytical Method Applied, and Discussion). SC-M and SS-C made substantial contribution to the design of the work. All four authors (RM, JL, SC-M, and SS-C) participated in several revisions during the process of creating, discussing, and improving the manuscript, with RM leading all revisions and guiding the continuous improvement of the manuscript; gave their consent that this final version is submitted for publication; and agreed in their co-responsibility regarding all aspects of the work, such as the accuracy of the data and the integrity of the research.

ACKNOWLEDGMENTS

REFERENCES

Abney, D. H., Warlaumont, A. S., Haussman, A., Ross, J. M., and Wallot, S. (2014). Using nonlinear methods to quantify changes in infant limb movements and vocalizations. *Front. Psychol.* 5:771. doi: 10.3389/fpsyg.2014.00771

Allison, D. B., and Gorman, B. S. (1993). Calculating effect sizes for meta-analysis: the case of the single case. *Behav. Res. Ther.* 31, 621–631. doi: 10.1016/0005-7967(93)90115-B

Anguera, M. T., Blanco-Villaseñor, Á, and Losada, J. L. (2001). Diseños observacionales, cuestión clave en el proceso de la Metodología Observacional. [Observational designs, a critical question in the process of Observational Methodology]. *Methodol. Behav. Sci.* 3, 135–160.

Apa Presidential Task Force on Evidence-Based Practice (2006). Evidence-based practice in psychology. *Am. Psychol.* 61, 271–285. doi: 10.1037/0003-066X.61.4.271

Bagner, D. M., Steinkopf, S. J., Miller-Loncar, C. L., Vohr, B. R., Hinckley, M., Eyberg, S. M., et al. (2009). Parent-Child Interaction Therapy for children born premature: a case study and illustration of vagal tone as a physiological measure of treatment outcome. *Cogn. Behav. Pract.* 16, 468–477. doi: 10.1016/j.cbpra.2009.05.002

Bakeman, R., and Quera, V. (2011). *Sequential Analysis and Observational Methods for the Behavioral Sciences.* Cambridge: Cambridge University Press.

Barlow, D. H., Nock, M. K., and Hersen, M. (Eds) (2009). *Single Case Experimental Designs: Strategies for Studying Behavior Change*, 3rd Edn. Boston, MA: Pearson.

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

Beretvas, S. N., and Chung, H. (2008). A review of meta-analyses of single-subject experimental designs: methodological issues and practice. *Evid. Based Commun. Assess. Interv.* 2, 129–141. doi: 10.1080/17489530802446302

Borckardt, J., and Nash, M. (2014). Simulation modelling analysis for small sets of single-subject data collected over time. *Neuropsychol. Rehabil.* 24, 492–506. doi: 10.1080/09602011.2014.895390

Borrego, J. Jr., Anhalt, K., Terao, S. Y., Vargas, E. C., and Urquiza, A. J. (2006). Parent-child interaction therapy with a Spanish-speaking family. *Cogn. Behav. Pract.* 13, 121–133. doi: 10.1016/j.cbpra.2005.09.001

Brossart, D. F., Parker, R. I., Olson, E. A., and Mahadevan, L. (2006). The relationship between visual analysis and five statistical analyses in a simple AB single-case research design. *Behav. Modif.* 30, 531–563. doi: 10.1177/0145445503261167

Brossart, D. F., Vannest, K., Davis, J., and Patience, M. (2014). Incorporating nonoverlap indices with visual analysis for quantifying intervention effectiveness in single-case experimental designs. *Neuropsychol. Rehabil.* 24, 464–491. doi: 10.1080/09602011.2013.868361

Bulté, I., and Onghena, P. (2012). When the truth hits you between the eyes: a software tool for the visual analysis of single-case experimental data. *Methodology* 8, 104–114. doi: 10.1027/1614-2241/a000042

Bulté, I., and Onghena, P. (2013). The single-case data analysis package: analysing single-case experiments with R software. *J. Mod. Appl. Stat. Methods* 12, 450–478.

Bunn, R., Burns, M. K., Hoffman, H. H., and Newman, C. L. (2005). Using incremental rehearsal to teach letter identification with a preschool-aged child. *J. Evid. Based Pract. Schools* 6, 124–134.

Busk, P. L., and Serlin, R. (1992). "Meta-analysis for single case research," in *Single-Case Research Design and Analysis: New Directions for Psychology and Education*, eds T. R. Kratochwill and J. R. Levin (Hillsdale, NJ: Lawrence Erlbaum), 187–212.

Busse, R. T., Kratochwill, T. R., and Elliott, S. N. (1995). Meta-analysis for single-case consultation outcomes: applications to research and practice. *J. Sch. Psychol.* 33, 269–285. doi: 10.1016/0022-4405(95)00014-D

Callahan, C. D., and Barisa, M. T. (2005). Statistical process control and rehabilitation outcome: the single-subject design reconsidered. *Rehabil. Psychol.* 50, 24–33. doi: 10.1037/0090-5550.50.1.24

Campbell, D. T., and Stanley, J. C. (1966). *Experimental and Quasi-experimental Designs for Research.* Chicago, IL: Rand McNally.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* 20, 37–46. doi: 10.1177/001316446002000104

Cohen, J. (1994). The earth is round (p < .05). *Am. Psychol.* 49, 997–1003. doi: 10.1037/0003-066X.49.12.997

Cordery, J. L., Morrisson, D., Wright, B. M., and Wall, T. B. (2010). The impact of autonomy and task uncertainty on team performance: a longitudinal field study. *J. Organ. Behav.* 31, 240–258. doi: 10.1002/job.657

Crane, D. R. (1985). Single-Case experimental designs in family therapy research: limitations and considerations. *Fam. Process* 24, 69–77. doi: 10.1111/j.1545-5300.1985.00069.x

Crosbie, J. (1987). The inability of the binomial test to control Type I error with single-subject data. *Behav. Assess.* 9, 141–150.

Danov, S. E., and Symons, F. J. (2008). A survey evaluation of the reliability of visual inspection and functional analysis graphs. *Behav. Modif.* 32, 828–839. doi: 10.1177/0145445508318606

Davis, D. H., Gagné, P., Fredrick, L. D., Alberto, P. A., Waugh, R. E., and Haardörfer, R. (2013). Augmenting visual analysis in single-case research with hierarchical linear modeling. *Behav. Modif.* 37, 62–89. doi: 10.1177/0145445512453734

DeCarlo, L. T., and Tryon, W. W. (1993). Estimating and testing correlation with small samples: a comparison of the C-statistic to modified estimator. *Behav. Res. Ther.* 31, 781–788. doi: 10.1016/0005-7967(93)90009-J

Edgington, E. S. (1980). Validity of randomization tests for one-subject experiments. *J. Educ. Stat.* 5, 235–251. doi: 10.3102/10769986005003235

Edgington, E. S., and Onghena, P. (2007). *Randomization Tests*, 4th Edn. London, UK: Chapman & Hall/CRC.

Eyberg, S. M., Nelson, M. M., and Boggs, S. R. (2008). Evidence-based psychosocial treatments for children and adolescents with disruptive behavior. *J. Clin. Child Adolesc. Psychol.* 37, 1–23. doi: 10.1080/15374410701820117

Eyberg, S. M., and Pincus, D. (1999). *Eyberg Child Behavior Inventory and Sutter-Eyberg Student Behavior Inventory: Professional Manual.* Odessa, FL: Psychological Assessment Resources.

Fabio, R. A., Castelli, I., Marchetti, A., and Antonietti, A. (2013). Training communication abilities in Rett Syndrome through reading and writing. *Front. Psychol.* 4:911. doi: 10.3389/fpsyg.2013.00911

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

Finn, M., and McDonald, S. (2014). A single case study of computerised cognitive training for older persons with mild cognitive impairment. *NeuroRehabilitation* 35, 261–270. doi: 10.3233/NRE-141121

Fisch, G. S. (2001). Evaluating data from behavioral analysis: visual inspection or statistical models? *Behav. Processes* 54, 137–154. doi: 10.1016/S0376-6357(01)00155-3

Fisher, W. W., Kelley, M. E., and Lomas, J. E. (2003). Visual aids and structured criteria for improving visual inspection and interpretation of single-case designs. *J. Appl. Behav. Anal.* 36, 387–406. doi: 10.1901/jaba.2003.36-387

Fisher, W. W., and Lerman, D. C. (2014). It has been said that, "There are three degrees of falsehoods: lies, damn lies, and statistics". *J. School Psychol.* 52, 243–248. doi: 10.1016/j.jsp.2014.01.001

Franklin, R. D., Gorman, B. S., Beasley, T. M., and Allison, D. B. (1996). "Graphical display and visual analysis," in *Design and Analysis of Single-Case Research*, eds R. D. Franklin, D. B. Allison, and B. S. Gorman (Mahwah, NJ: Lawrence Erlbaum), 119–158.

Gage, N. A., and Lewis, T. J. (2014). Hierarchical linear modeling meta-analysis of single-subject design research. *J. Spec. Educ.* 48, 3–16. doi: 10.1177/0022466912443894

Gast, D. L., and Spriggs, A. D. (2010). "Visual analysis of graphic data," in *Single Subject Research Methodology in Behavioral Sciences*, ed. D. L. Gast (London, UK: Routledge), 199–233.

Gibson, G., and Ottenbacher, K. (1988). Characteristics influencing the visual analysis of single-subject data: an empirical analysis. *J. Appl. Behav. Sci.* 24, 298–314. doi: 10.1177/0021886388243007

Glass, G. V., McGaw, B., and Smith, M. L. (1981). *Meta-analysis in Social Research.* Beverly Hills, CA: Sage.

Grissom, R. J. (1994). Probability of the superior outcome of one treatment over another. *J. Appl. Psychol.* 79, 314–316. doi: 10.1037/0021-9010.79.2.314

Grissom, R. J., and Kim, J. J. (2012). *Effect Size for Research: Univariate and Multivariate Applications*, 2nd Edn. London, UK: Routledge.

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

Harrop, J. W., and Velicer, W. F. (1985). A comparison of three alternative methods of time series model identification. *Multivariate Behav. Res.* 20, 27–44. doi: 10.1207/s15327906mbr2001_2

Heyvaert, M., and Onghena, P. (2014). Randomization tests for single-case experiments: state of the art, state of the science, and state of the application. *J. Contextual Behav. Sci.* 3, 51–64. doi: 10.1016/j.jcbs.2013.10.002

Heyvaert, M., Wendt, O., Van Den Noortgate, W., and Onghena, P. (2015). Randomization and data-analysis items in quality standards for single-case experimental studies. *J. Spec. Educ.* 49, 146–156. doi: 10.1177/0022466914525239

Houle, T. T. (2009). "Statistical analyses for single-case experimental designs," in *Single Case Experimental Designs: Strategies for Studying Behavior Change*, 3rd Edn, eds D. H. Barlow, M. K. Nock, and M. Hersen (Boston, MA: Pearson), 271–305.

Howick, J., Chalmers, I., Glasziou, P., Greenhalgh, T., Heneghan, C., Liberati, A., et al. (2011). *The 2011 Oxford CEBM Evidence Table (Introductory Document).* Oxford: Centre for Evidence-Based Medicine.

Hugdahl, K., and Ǒst, L.-G. (1981). On the difference between statistical and clinical significance. *Behav. Assess.* 3, 289–295.

Huitema, B. E., McKean, J. W., and Laraway, S. (2007). Time series intervention analysis using ITSACORR: fatal flaws. *J. Mod. Appl. Stat. Methods* 6, 367–379.

Institute of Education Sciences (2013). *Request for Applications: Statistical and Research Methodology in Education.* Available at: http://ies.ed.gov/funding/pdf/2014_84305D.pdf

Jacobson, N. S., and Truax, P. (1991). Clinical significance: a statistical approach to meaningful change in psychotherapy research. *J. Consult. Clin. Psychol.* 59, 12–19. doi: 10.1037/0022-006X.59.1.12

Jacoby, W. G. (2000). Loess: a nonparametric, graphical tool for depicting relationships between variables. *Electoral Stud.* 19, 577–613. doi: 10.1016/S0261-3794(99)00028-1

Jamieson, M., Cullen, B., McGee-Lennon, M., Brewster, S., and Evans, J. J. (2014). The efficacy of cognitive prosthetic technology for people with memory impairments: a systematic review and meta-analysis. *Neuropsychol. Rehabil.* 24, 419–444. doi: 10.1080/09602011.2013.825632

Jenson, W. R., Clark, E., Kircher, J. C., and Kristjansson, S. D. (2007). Statistical reform: evidence-based practice, meta-analyses, and single subject designs. *Psychol. Schools* 44, 483–493. doi: 10.1002/pits.20240

Johnston, J. M., and Pennypacker, H. S. (2008). *Strategies and Tactics of Behavioral Research*, 3rd Edn. New York, NY: Routledge.

Jonsson, G. K., Anguera, M. T., Blanco-Villaseñor, Á, Losada, J. L., Hernández-Mendo, A., Ardá, T., et al. (2006). Hidden patterns of play interaction in soccer using SOF-CODER. *Behav. Res. Methods* 38, 372–381. doi: 10.3758/BF03192790

Kazdin, A. E. (1978). Methodological and interpretive problems of single-case experimental designs. *J. Consult. Clin. Psychol.* 46, 629–642. doi: 10.1037/0022-006X.46.4.629

Kazdin, A. E. (1982). *Single-Case Research Designs: Methods for Clinical and Applied Settings.* New York, NY: Oxford University Press.

Kazdin, A. E. (1984). *Behavior Modification in Applied Settings*, 3rd Edn. Homewood, IL: The Dorsey Press.

Kazdin, A. E. (1999). The meanings and measurements of clinical significance. *J. Consult. Clin. Psychol.* 67, 332–339. doi: 10.1037/0022-006X.67.3.332

Kendall, P. C. (1999). Clinical significance. *J. Consult. Clin. Psychol.* 67, 283–284. doi: 10.1037/0022-006X.67.3.283

Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., et al. (2010). *Single Case Designs Technical Documentation. In What Works Clearinghouse: Procedures and Standards Handbook (Version 2.0).* Available at: http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf

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

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

Kratochwill, T. R., Stoiber, K. C., and Gutkin, T. B. (2001). Empirically supported interventions in school psychology: the role of negative results in outcome research. *Psychol. Schools* 37, 399–413. doi: 10.1177/0741932512452794

Lane, J. D., and Gast, D. L. (2014). Visual analysis in single case experimental design studies: brief review and guidelines. *Neuropsychol. Rehabil.* 24, 445–463. doi: 10.1080/09602011.2013.815636

Ledford, J., and Gast, D. L. (2014). Measuring procedural fidelity in behavioural research. *Neuropsychol. Rehabil.* 24, 332–348. doi: 10.1080/09602011.2013.861352

Levin, J. R., Evmenova, A. S., and Gafurov, B. S. (2014). "The single-case data-analysis ExPRT (Excel Package of Randomization Tests)," in *Single-Case Intervention Research: Methodological and Statistical Advances*, eds T. R. Kratochwill and J. R. Levin (Washington, DC: American Psychological Association), 185–219.

Levin, J. R., Ferron, J. M., and Kratochwill, T. R. (2012). Nonparametric statistical tests for single-case systematic and randomized ABAB...AB and alternating treatment intervention designs: new developments, new directions. *J. Sch. Psychol.* 50, 599–624. doi: 10.1016/j.jsp.2012.05.001

Lieberman, R. G., Yoder, P. J., Reichow, B., and Wolery, M. (2010). Visual analysis of multiple baseline across participants graphs when change is delayed. *Sch. Psychol. Q.* 25, 28–44. doi: 10.1037/a0018600

Maggin, D. M., Briesch, A. M., and 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 Spec. Educ.* 34, 44–58. doi: 10.1177/0741932511435176

Maggin, D. M., Briesch, A. M., Chafouleas, S. M., Ferguson, T. D., and Clark, C. (2014). A comparison of rubrics for identifying empirically supported practices with single-case research. *J. Behav. Educ.* 23, 287–311. doi: 10.1007/s10864-013-9187-z

Manolov, R., and Rochat, L. (2015). Further developments in summarising and meta-analysing single-case data: an illustration with neurobehavioural interventions in acquired brain injury. *Neuropsychol. Rehabil.* 25, 637–662. doi: 10.1080/09602011.2015.1064452

Manolov, R., Sierra, V., Solanas, A., and Botella, J. (2014). Assessing functional relations in single-case designs: quantitative proposals in the context of the evidence-based movement. *Behav. Modif.* 38, 878–913. doi: 10.1177/0145445514545679

Manolov, R., and Solanas, A. (2008). Comparing N = 1 effect size indices in presence of autocorrelation. *Behav. Modif.* 32, 860–875. doi: 10.1177/0145445508318866

Manolov, R., Solanas, A., Bulté, I., and Onghena, P. (2010). Data-division-specific robustness and power for ABAB designs. *J. Exp. Educ.* 78, 191–214. doi: 10.1080/00220970903292827

Manolov, R., Solanas, A., Sierra, V., and Evans, J. J. (2011). Choosing among techniques for quantifying single-case intervention effectiveness. *Behav. Ther.* 42, 533–545. doi: 10.1016/j.beth.2010.12.003

Matyas, T. A., and Greenwood, K. M. (1996). "Serial dependency in single-case time series," in *Design and Analysis of Single-Case Research*, eds R. D. Franklin, D. B. Allison, and B. S. Gorman (Mahwah, NJ: Lawrence Erlbaum), 215–243.

McGraw, K. O., and Wong, S. P. (1992). A common language effect size statistic. *Psychol. Bull.* 111, 361–365. doi: 10.1037/0033-2909.111.2.361

McKnight, S. D., McKean, J. W., and Huitema, B. E. (2000). A double bootstrap method to analyze linear models with autoregressive error terms. *Psychol. Methods* 3, 87–101. doi: 10.1037/1082-989X.5.1.87

Michael, J. (1974). Statistical inference for individual organism research: mixed blessing or curse? *J. Appl. Behav. Anal.* 7, 647–653. doi: 10.1901/jaba.1974.7-647

Miller, M. J. (1985). Analyzing client change graphically. *J. Couns. Dev.* 63, 491–494. doi: 10.1002/j.1556-6676.1985.tb02743.x

Moeyaert, M., Ferron, J., Beretvas, S., and Van Den Noortgate, W. (2014a). From a single-level analysis to a multilevel analysis of since-case experimental designs. *J. Sch. Psychol.* 52, 191–211. doi: 10.1016/j.jsp.2013.11.003

Moeyaert, M., Ugille, M., Ferron, J., Beretvas, S. N., and Van den Noortgate, W. (2014b). The influence of the design matrix on treatment effect estimates in the quantitative analyses of single-case experimental designs research. *Behav. Modif.* 38, 665–704. doi: 10.1177/0145445514535243

Ninci, J., Vannest, K. J., Willson, V., and Zhang, N. (2015). Interrater agreement between visual analysts of single-case data: a meta-analysis. *Behav. Modif.* 39, 510–541. doi: 10.1177/0145445515581327

Odom, S. L., Brantlinger, E., Gersten, R., Horner, R. H., Thompson, B., and Harris, K. R. (2005). Research in special education: scientific methods and evidence-based practices. *Except. Child.* 71, 137–148. doi: 10.1177/001440290507100201

O'Neill, B., Best, C., Gillespie, A., and O'Neill, L. (2013). Automated prompting technologies in rehabilitation and at home. *Soc. Care Neurodisability* 4, 17–28. doi: 10.1108/20420911311302281

Ottenbacher, K. J. (1990). Visual inspection of single-subject data: an empirical analysis. *Ment. Retard.* 28, 283–290.

Parker, R. I. (2006). Increased reliability for single-case research results: is bootstrap the answer? *Behav. Ther.* 37, 326–338. doi: 10.1016/j.beth.2006.01.007

Parker, R. I., and Brossart, D. F. (2003). Evaluating single-case research data: a comparison of seven statistical methods. *Behav. Ther.* 34, 189–211. doi: 10.1016/S0005-7894(03)980013-8

Parker, R. I., Cryer, J., and Byrns, G. (2006). Controlling baseline trend in single-case research. *Sch. Psychol. Q.* 21, 418–443. doi: 10.1037/h0084131

Parker, R. I., and Vannest, K. J. (2009). An improved effect size for single-case research: nonoverlap of all pairs. *Behav. Ther.* 40, 357–367. doi: 10.1016/j.beth.2008.10.006

Parker, R. I., Vannest, K. J., and Davis, J. L. (2011a). Effect size in single-case research: a review of nine nonoverlap techniques. *Behav. Modif.* 35, 303–322. doi: 10.1177/0145445511399147

Parker, R. I., Vannest, K. J., Davis, J. L., and Sauber, S. B. (2011b). Combining nonoverlap and trend for single-case research: Tau-U. *Behav. Ther.* 42, 284–299. doi: 10.1016/j.beth.2010.08.006

Pashler, H., and Wagenmakers, E. J. (2012). Editors' introduction to the special section on replicability in psychological science: a crisis of confidence? *Perspect. Psychol. Sci.* 7, 528–530. doi: 10.1177/1745691612465253

Pazzagli, C., Laghezza, L., Manaresi, F., Mazzeschi, C., and Powell, B. (2014). The circle of security parenting and parental conflict: a single case study. *Front. Psychol.* 5:887. doi: 10.3389/fpsyg.2014.00887

R Core Team (2013). *R: A Language and Environment for Statistical Computing.* Vienna: R Foundation for Statistical Computing.

Robey, R. R., Schultz, M. C., Crawford, A. B., and Sinner, C. A. (1999). Single-subject clinical outcome research: designs, data, effect sizes, and analysis. *Aphasiology* 13, 445–473. doi: 10.1080/026870399402028

Scruggs, T. E., and Mastropieri, M. A. (2013). PND at 25: past, present, and future trends in summarizing single-subject research. *Remedial Spec. Educ.* 34, 9–19. doi: 10.1177/0741932512440730

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

Sidman, M. (1960). *Tactics of Scientific Research: Evaluating Experimental Data in Psychology.* New York, NY: Basic Books.

Simonton, D. K. (1977). Cross-sectional time-series experiments: some suggested statistical analyses. *Psychol. Bull.* 84, 489–502. doi: 10.1037/0033-2909.84.3.489

Smith, J. D. (2012). Single-Case experimental designs: a systematic review of published research and current standards. *Psychol. Methods* 17, 510–550. doi: 10.1037/a0029312

Solanas, A., Manolov, R., and Onghena, P. (2010a). Estimating slope and level change in N = 1 designs. *Behav. Modif.* 34, 195–218. doi: 10.1177/0145445510363306

Solanas, A., Manolov, R., and Sierra, V. (2010b). Lag-one autocorrelation in short series: estimation and hypothesis testing. *Psicológica* 31, 357–381.

Solmi, F., Onghena, P., Salmaso, L., and Bulté, I. (2014). A permutation solution to test for treatment effects in alternation design single-case experiments. *Commun. Stat.* 43, 1094–1111. doi: 10.1080/03610918.2012.725295

Solomon, B. G., Howard, T. K., and Stein, B. L. (2015). Critical assumptions and distribution features pertaining to contemporary single-case effect sizes. *J. Behav. Educ.* 24, 438–458. doi: 10.1007/s10864-015-9221-4

Suen, H. K., and Ary, D. (1989). *Analyzing Quantitative Behavioral Data.* Hillsdale, NJ: Lawrence Erlbaum.

Swaminathan, H., Rogers, H. J., and Horner, R. H. (2014). An effect size measure and Bayesian analysis of single-case designs. *J. Sch. Psychol.* 52, 213–230. doi: 10.1016/j.jsp.2013.12.002

Tate, R. L., Perdices, M., Rosenkoetter, U., Shadish, W., Vohra, S., and Wilson, B. (in press). The single-case Reporting guideline in behavioural interventions (SCRIBE) 2015 statement. *Arch. Sci. Psychol.*

Tate, R. L., Perdices, M., Rosenkoetter, U., Wakima, D., Godbee, K., Togher, L., et al. (2013). Revision of a method quality rating scale for single-case experimental designs and n-of-1 trials: the 15-item Risk of Bias in N-of-1 Trials (RoBiNT) Scale. *Neuropsychol. Rehabil.* 23, 619–638. doi: 10.1080/09602011.2013.824383

Taylor, L. K., and Weems, C. F. (2011). Cognitive-behavior therapy for disaster-exposed youth with posttraumatic stress: results from a multiple-baseline examination. *Behav. Ther.* 42, 349–363. doi: 10.1016/j.beth.2010.09.001

Theiler, J., Eubank, S., Longtin, A., Galdrikian, B., and Farmer, J. D. (1992). Testing for nonlinearity in time series: the method of surrogate data. *Physica D* 58, 77–94. doi: 10.1016/0167-2789(92)90102-S

Tryon, W. W. (1982). A simplified time-series analysis for evaluating treatment interventions. *J. Appl. Behav. Anal.* 15, 423–429. doi: 10.1901/jaba.1982.15-423

Tukey, J. W. (1977). *Exploratory Data Analysis.* London, UK: Addison-Wesley.

Vannest, K. J., Parker, R. I., and Gonen, O. (2011). *Single Case Research: Web Based Calculators for SCR Analysis. (Version 1.0) [Web-Based Application]*. College Station, TX: Texas A&M University.

Winkens, I., Ponds, R., Pouwels-van den Nieuwenhof, C., Eilander, H., and van Heugten, C. (2014). Using single-case experimental design methodology to evaluate the effects of the ABC method for nursing staff on verbal aggressive behaviour after acquired brain injury. *Neuropsychol. Rehabil.* 24, 349–364. doi: 10.1080/09602011.2014.901229

Wolery, M., Busick, M., Reichow, B., and Barton, E. E. (2010). Comparison of overlap methods for quantitatively synthesizing single-subject data. *J. Spec. Educ.* 44, 18–29. doi: 10.1177/0022466908328009

Ximenes, V. M., Manolov, R., Solanas, A., and Quera, V. (2009). Factors affecting visual inference in single-case designs. *Span. J. Psychol.* 12, 823–832. doi: 10.1017/S1138741600002195

Young, L. C. (1941). On the randomness in ordered sequences. *Ann. Math. Stat.* 12, 293–300. doi: 10.1214/aoms/1177731711