Journal of the Experimental Analysis of Behavior 2022, **118,** 462–492 NUMBER 3 (NOVEMBER)

Defining and assessing immediacy in single-case experimental designs

Rumen Manolov[1] and Patrick Onghena[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

Immediacy is one of six data aspects (alongside level, trend, variability, overlap, and consistency) that has to be accounted for when visually analyzing single-case data. Given that it is one of the aspects that has received considerably less attention than other data aspects, the current text offers a review of the proposed conceptual definitions of immediacy (i.e., what it refers to) and also of the suggested operational definitions (i.e., how exactly is it assessed and/or quantified). Provided that a variety of conceptual and operational definitions is identified, we propose following a sensitivity analysis using a randomization test for assessing immediate effects in single-case experimental designs, by identifying when changes were most clear. In such a sensitivity analysis, the immediate effects are tested for multiple possible intervention points and for different possible operational definitions. Robust immediate effects can be detected if the results for the different operational definitions converge.

Key words: single-case experimental designs, immediacy, latency, randomization

When performing a visual analysis of single-case experimental designs (SCED) data, six data aspects are usually assessed: level, trend, variability, immediacy, overlap, and consistency (Kratochwill et al., 2013; What Works Clearinghouse, 2017). This recommendation is strongly based on the work by Horner et al. (2005) and has been echoed several times across multiple publications presenting SCEDs across a variety of fields such as health psychology (Epstein et al., 2021), special education (Ledford et al., 2019; Maggin et al., 2018), rehabilitation (Graham et al., 2012), neurology (Lobo et al., 2017), and behavior modification (Ninci, 2019).

Some of these six data aspects have received greater attention in the scientific literature than others. Specifically, the data aspect "overlap" has received a great deal of attention, with the Percentage of Nonoverlapping Data (Scruggs et al., 1987) as one of the first and most widely adopted quantification of effect size in SCEDs, and with other more recent nonoverlap indices as the most frequently used alternatives for the quantification of SCED effects (Jamshidi et al., 2022; Maggin, O'Keefe, & Johnson, 2011; Radley et al., 2020). The different nonoverlap indices have also been compared multiple times (e.g., Lenz, 2013; Parker, Vannest, & Davis, 2011; Rakap, 2015), and included in studies both with real data (e.g., M. Chen et al., 2016; Wolery et al., 2010) and with generated data (e.g., Giannakakos & Lanovaz, 2019; Tarlow, 2017).

Another set of common quantifications focuses on level, such as the within-case standardized mean difference (Busk & Serlin, 1992), or the between-case standardized mean difference (Hedges et al., 2012, 2013). There is evidence that level is the most commonly quantified data aspect (Tanious & Onghena, 2021), with the within-case standardized mean difference as a common measure at least in some contexts (Radley et al., 2020). Regarding the between-case standardized mean difference, it has been an object of both technical reports (Shadish et al., 2015), a recommendation in Version 4.1 of the What Works Clearinghouse (2020) standards, and several illustrations (e.g., Barton et al., 2017; Maggin et al., 2017). Beyond expressing the mean difference in standard

Address correspondence to: Rumen Manolov, Department de Psicologia Social i Psicologia Quantitativa, Universitat de Barcelona, Passeig de la Vall d'Hebron 171, 08035, Barcelona, Spain. Phone: +34 934025072; E-mail: rrumenovl3@ub.edu

doi: 10.1002/jeab.799

deviation units, a comparison in terms of level is also performed via the log-response ratio, which can be expressed as percentage change (Pustejovsky, 2018), or via the mean baseline reduction (Olive & Smith, 2005).

The possibilities for assessing trends have also been discussed in several articles, including a focus on the split-middle technique (Fisher et al., 2003; Miller, 1985), the trisplit technique (Parker et al., 2014) or Tukey's resistant trend line (Franklin, Gorman, et al., 1996), the Mann-Kendall test (Hamed & Rao, 1998), ordinary least squares estimation (Moeyaert, Ugille, et al., 2014; Parker et al., 2006) or generalized least squares estimation (Swaminathan et al., 2014), Theil-Sen resistant trend (Tarlow & Brossart, 2018; Vannest et al., 2012), monotonic trend (Parker, Vannest, Davis, & Sauber, 2011), and generalized additive models (Sullivan et al., 2015). There have also been texts dedicated to discussing several approaches for fitting trend lines (L.-T. Chen et al., 2019; Manolov, 2018).

Another aspect that recently received attention is the assessment of consistency, focusing on the consistency of data patterns in similar phases and also on the consistency of effects (Tanious, De, Michiels, et al., 2019; Tanious et al., 2020, 2021). Other proposals for assessing consistency have focused on specific designs (see Manolov et al., 2021, dealing with alternation designs), or specific graphical representations (Manolov & Tanious, 2022). Moreover, there have been proposals for assessing consistency of effects in the context of multilevel models (Manolov & Ferron, 2020).

The remaining two data aspects, variability and immediacy, have received less attention in terms of specific quantifications for making their assessment more objective. Most proposals referring to variability are based on or understood as visual aids. Moreover, most proposals also refer to the stability of the baseline. On the one hand, the stability envelope focuses on the within-phase data pattern with respect to variability and trend stability (Lane & Gast, 2014, see also Swan et al., 2020, for other operational definitions). On the other hand, there have been proposals for extending the baseline level with a variability band as in statistical process control (Callahan & Barisa, 2005; Pfadt & Wheeler, 1995) or for extending the baseline trend with a variability band (Manolov & Vannest, 2019). In both of these latter proposals, the focus is placed on the comparison between the projected baseline and the actual intervention phase data, rather than variability per se. Similarly, variability has been discussed in the context of standardized mean differences, for which variability is relevant for standardizing the difference; a distinction between within-subject and between-subject variability is crucial for the interpretation of standardized mean differences (Beretvas & Chung, 2008; Odom et al., 2018). In that sense, the importance of variability in these latter quantifications stems from the fact that differences in level cannot be assessed without taking variability within phases into account. Finally, "changes in variability between phases" (as a research question, as a prediction or effect size measure) is very uncommon.

Finally, the assessment of immediacy has not seen many specific developments or any broad reviews. As later sections will show, the two exceptions are a recommendation from the What Works Clearinghouse (2017) and a proposal based on Bayesian analysis (Natesan & Hedges, 2017; Natesan Batley, Minka, & Hedges, 2020; Natesan Batley et al., 2021). There are also several questions remaining to be answered, related to the exact number of measurements that have to be included when assessing immediacy, the data features that are object of this assessment, or even in terms of how best to translate the conceptual definitions of immediacy into operational definitions.

Aim and Organization of the Text

Considering the interest of single-case researchers in verifying a functional relation between variables and the potential relevance of immediacy for verification of such a functional relation, defining "immediacy" is an important endeavor. Thus, the initial aim of the current text is to review some conceptual and operational definitions that we could find in the SCED literature for what an immediate effect is and how it should be assessed. In order to meet this aim, an initial section presents several conceptual definitions found in textbooks and articles, while also discussing the importance of immediacy for the assessment of experimental control, for choosing a specific design, and for quantitative data analysis. Afterwards, we review the way in which immediate effects have been recommended to be assessed, in terms of how many data points need to be taken in consideration and which data features (e.g., level, trend, overlap) are the object of the assessment.

Given the vagueness and ambiguity in conceptual and operational definitions that we could identify in the literature, we focus on a proposal for an operational definition that could answer a twofold question: (a) for which potential moment of change in phase is the difference largest (i.e., the evidence for a behavioral change strongest); and (b) is this evidence well-aligned with the concept of an "immediate effect" or with the kind of effect expected. This proposal entails a combination of randomization test logic (Edgington, 1967, 1996) and a sensitivity analysis (which could be understood as a way of following the multiverse approach advocated by Steegen et al., 2016).

The proposal is illustrated with real data exhibiting different data patterns, but because of the limited number of examples (and the multitude of possible designs and data patterns) the current study is best viewed as a proof-of-concept, rather than full-scale empirical validation.

Immediacy: Conceptual Definitions and Importance

Search Procedure

In order to identify conceptual and operational definitions of immediacy, the search process included the following steps. First, we searched peer-reviewed articles in the Web of Science and PsycINFO databases with the terms ("single-case" OR "single-subject") AND "immedia*". Second, we read the key articles that prompted us to study immediacy (i.e., Horner & Kratochwill, 2012; Natesan & Hedges, 2017; Swan & Pustejovsky, 2018; Verboon & Peters, 2020) for further references cited in them or for other articles citing them. Third, we checked all textbooks on SCED methodology that we know of (i.e., Barker et al., 2011; Barlow et al., 2009; Franklin, Allison, & Gorman, 1996; Janosky et al., 2009; Kazdin, 2020; Kennedy, 2005; Kratochwill & Levin, 2014; Ledford & Gast, 2018; Morgan & Morgan, 2009; Morley, 2018; Poling & Fuqua, 1986; Riley-Tillman et al., 2020; Sidman, 1960; Tate & Perdices, 2019; van de Schoot & Miočević, 2020) for text on immediacy.

Conceptual Definitions

What is Immediacy?

The concept of immediacy appears inherently connected to the concept of latency, especially if we are looking for a definition of "immediacy" that is not tautological. Specifically, "latency of change refers to the amount of time for the intervention to have an impact on the behavior. Intervention effects can be immediate or delayed" (Riley-Tillman et al., 2020, p. 80). Analogously, Kazdin (2020) defines latency as the period between the onset of one condition and the change in performance, with brief, rapid or short latency referring to an immediate effect. Another term used to refer to immediate effects is "abrupt" (Parsonson & Baer, 1978). Using this term, Maggin et al. (2018) offer one of their definitions of immediate effect: "large and abrupt change in the data corresponding to researchers introducing or withdrawing the intervention" (p.188). These authors also include another definition: "magnitude and quickness of change in the data corresponding to change in the intervention" (Maggin et al., 2018, p. 189). Several other definitions also reflect the idea of a degree to which the effect is immediate. For instance, Ledford et al. (2018, p. 10) refer to the "extent to which data change simultaneously with a condition change (Ledford et al., 2018, p.10)." Barton et al. (2018, p. 191) state that the immediacy of change is "the degree to which behavior change occurs as soon as the intervention is introduced". Similarly, Kennedy (2005, p. 203) states that the immediacy of effect or rapidity of change refers to "how quickly a change in the data pattern is produced after the phase change." Equivalently, Horner and Odom (2014, p. 34) refer to "how quickly is change demonstrated." Thus, these latter definitions refer to degrees and not to a dichotomous decision regarding whether the effect is immediate or not. Finally, Morley (2018, p. 115) refers to "point of change" when assessing the moment in which the behavior change occurs and whether the intervention acts rapidly.

What is the Focus of the Assessment of Immediacy?

According to one of the definitions presented in Version 4.0 of the Standards, "immediacy of the effect refers to the change in level between the last three data points in one phase and the first three data points of the next" (What Works Clearinghouse, 2017, p. A-7). Similarly, an emphasis on the change in level is present in other sources (Kilgus

et al., 2016; Ledford et al., 2019; Ninci, 2019; Tate & Perdices, 2019; Wolfe et al., 2019). Kazdin (2020, p. 357) also refers to "shift in level when a phase is changed" when referring to what happened immediately after the intervention was introduced or withdrawn.

In contrast, later in the text of Version 4.0 of the Standards, "immediacy of the effect compares the extent to which the level, trend, and variability of the last three data points in one phase are distinguishably different from the first three data points in the next" (What Works Clearinghouse, 2017, p. A-11). Similar to this broader definition, several authors (Barton et al., 2019; Gage & Lewis, 2013; Haegele & Hodge, 2015; Horner & Odom, 2014; Kennedy, 2005; Spear et al., 2013) mention all three data aspects: level, trend, variability when referring to immediacy. In contrast, Lane et al. (2021) only mention level and trend. Similarly, Levin et al. (2021) refer to two kinds of immediate effects: an immediate abrupt effect (change in level) and an immediate gradual effect (change in trend). Finally, Tate et al. (2014) also include overlap, apart from level, trend, and variability when defining the immediate effect. Similar to Tate et al., Morley (2018) mentions specifically overlap when assessing the point of change.

The apparent lack of consistency in the literature in relation to the focus of the assessment of immediacy suggests that dealing with how immediate effects are defined is a relevant topic.

What Is the Opposite of Immediacy?

The previously mentioned distinction between an immediate abrupt and an immediate gradual effect (Levin et al., 2021) is relevant for defining the opposite of an immediate effect. Specifically, Levin et al. (2021) refer to delayed effects, as effects for which the change in level or in trend starts later in the intervention phase. Similarly, Houle (2009) illustrates both abrupt and gradual effects taking place immediately with the change in phase. This also agrees with Tate and Perdices (2019) and Riley-Tillman et al. (2020), who oppose immediate to delayed effects.

A somewhat less clear distinction is present in Natesan and Hedges (2017) and Natesan Batley, Minka, & Hedges (2020), who refer to situations in which an immediate effect is not likely, by using several terms such as latency, gradual effects, and delayed effects. In another text by these authors (Natesan Batley et al., 2021), they do stress delay as an opposite term of immediacy, and refer to the need to model slopes when working with gradual effects. Another apparent mixing of terms is present in Swan and Pustejovsky (2018) who present a "gradual effects model", which incorporates in this formulaic expression a parameter that determines the delay in reaching the full effect of the treatment. However, this model refers to the delay in reaching the asymptote, not the delay in onset of the treatment effect, as all effects depicted in Swan and Pustejovsky start immediately.

In summary, it can be concluded that when the effect (abrupt or gradual) is not immediate, this entails a delay or latency. Therefore, "immediate effects" can be expressed in terms of the remaining data aspects (level, trend, variability, and overlap).

Two other aspects are worth mentioning. On the one hand, the amount of latency can be conceptualized as continuous, distinguishing between different degrees of (short and long) latency. On the other hand, the amount of latency is a separate issue from the assessment of whether the effect is transitory (temporary) or permanent and also a separate issue from the abrupt or gradual nature of the effect.

Importance of Immediacy for Inferring Causal Relations

It has been acknowledged that one of the relevant aspects of the data pattern, when assessing experimental control or the presence of a functional relation, is whether the effect of the intervention is immediate (Cook et al., 2015; Horner et al., 2005). Specifically, immediate effects are more easily and more confidently interpreted as being due to the intervention and not to an external factor (Barton et al., 2018; L. L. Cohen et al., 2014; Kennedy, 2005; Ledford, 2018; Riley-Tillman et al., 2020; Tankersley et al., 2006; What Works Clearinghouse, 2017). Conversely, "history" is considered to be a greater threat for internal validity, in case the behavioral change does not take place immediately after the treatment onset (Petursdottir & Carr, 2018).

However, when using a SCED, the effects are not necessarily always immediate (Dallery & Raiff, 2014; Maggin et al., 2018; Wolfe et al., 2019), for instance when studying academic or developmental skills (Kratochwill et al., 2014;

Lieberman et al., 2010) or in rehabilitation (Krasny-Pacini & Evans, 2018). Other examples of expected lack of immediacy include transition states (Brogan et al., 2019) and extinction bursts, that is, data in the beginning of the intervention phase that are worse than the preceding baseline measurements, when a reinforcer is removed (Barnard-Brak et al., 2020; Riley-Tillman et al., 2020). Moreover, in some cases distressing interventions may cause immediate distress, but not the expected positive immediate effect (Holman & Koerner, 2014). Therefore, it has been suggested that "it is prudent not to be under the exclusive control of abrupt changes in trend across adjacent phases, or abrupt changes in level between adjacent phases" (Parsonson & Baer, 1986, p. 172). This entails assigning less weight to immediacy when visually inspecting the data using the six commonly assessed data features (Horner & Kratochwill, 2012; Wolfe et al., 2019). Two very relevant aspects when judging whether an effect that is not immediate is trustworthy is whether there was an expectation for a delayed effect (Barton et al., 2018; What Works Clearinghouse, 2017) and whether the latency of change is consistent across replications (Ledford et al., 2018; Lieberman et al., 2010). Moreover, Maggin et al. (2018) emphasize the importance of trend and the magnitude of change over time when an immediate effect cannot be expected, and Barton et al. (2018) emphasize the consistency of the magnitude of change in level or trend replications. Finally, in relation to consistency, it is relevant to explore, in the context of the same study, why the effect for certain participants is immediate and for others not, if such a result happens to be obtained (Kipfmiller et al., 2019).

Finally, immediacy has also been commented on in relation to social validity, by posing the question regarding whether a powerful effect is an immediate and dramatic change that requires a lot of resources to be maintained or a gradual and self-maintained effect that requires few resources (Franklin, Gorman, et al., 1996).

In summary, immediacy is theoretically important for establishing a causal relation, but in many behavioral applications the effects of the intervention and the behavioral change are not likely to be immediate. In those applications, the assessment of immediacy has to be replaced by the assessment of whether the actually observed effect matches the effect that was predicted and expected on the basis of previous research and the expert knowledge of the behavioral process studied. Furthermore, maintenance of any effect observed (immediate or gradual) is a relevant aspect of the practical significance of any intervention.

Importance of Immediacy for Choosing a Design

Whether the effect can be expected to be immediate or not also has an effect on the kind of SCED to be used. On the one hand, when the effect is expected to be delayed and not immediate a phase design (e.g., reversal design) is preferable over an alternation design. On the other hand, if the desired final level of the target behavior cannot be achieved immediately and abruptly, but in a more gradual fashion, Lane et al. (2017) and Tate and Perdices (2019) recommend using changing criterion designs. However, in the context of this design, when comparing the behavior level to the prespecified criterion, an immediate adherence to this level is also required (Ledford et al., 2019), whereas McDougall et al. (2006) state that a minimal latency of change is also acceptable.

In contrast to changing criterion designs, when immediate effects are expected, an alternating treatments design can be used (Holcombe et al., 1994). It is also required for the effect to disappear immediately when an intervention is withdrawn or when conditions switch in order to avoid carryover effects. Another option is to include washout periods by design (Kwasnicka et al., 2019).

For a concurrent multiple-baseline designs, a delayed effect may distort the assessment of the verification period (Carr, 2005), in which the intervention is already introduced for some tiers but not yet for others. Thus, discarding history as an alternative explanation for the changes in the performance would need to be based on observing a consistent pattern across tiers.

Finally, in relation to phase designs (multiple-baseline and ABAB), immediate effects allow for briefer phases (Kennedy, 2005), when a SCED follows the principles of response-guided experimentation.

Importance of Immediacy for Quantitative Data Analysis

Immediate effects are assumed when using some data analytical procedures such as the between-case standardized mean difference

(Shadish et al., 2014; also referred to as "design-comparable effect size" and recommended by What Works Clearinghouse, 2020). In contrast, another quantification, expressed as percentage change (Hershberger et al., 1999; Olive & Smith, 2005) from the baseline focuses only on the last three data points from each phase and thus does not allow for quantifying immediate effects. A third kind of data analytical procedure allows estimating when the largest rate in change takes place via a logistic model (Verboon & Peters, 2020). This way of analyzing the data assumes that the change is progressive, which is similar to the data analytical approach in the gradual effects model by Swan and Pustejovsky (2018).

In the light of the data analytical options mentioned so far, the choice of how to analyze SCED data can be based on the researchers' expectations. In a fourth data-analytical approach, randomization tests, the test statistic is precisely chosen according to the expected effect and, thus, it can either focus on quantifying an immediate effect (Michiels & Onghena, 2019) or a delayed effect (Levin et al., 2017).

A fifth option arises in regression-based analysis of SCED data. For instance, a model based on generalized least squares regression entails comparing the projected baseline trend with the fitted intervention phase trend (Maggin, Swaminathan, et al., 2011). This model allows specifying for which moment in time (i.e., for which intervention phase measurement occasion) to compare the expected baseline and intervention levels, although its authors advocate for another option: namely, to compute an average difference that combines level and trend (Swaminathan et al., 2014). In contrast, the results of another regression-based approach, using a piecewise model, are usually expressed as separate quantifications of change in slope and change in level. Regarding the quantification of the change in level, it can reflect an immediate effect for the first intervention phase measurement occasion or it can focus on a different moment in time (Moeyaert, Ugille et al., 2014).

Finally, instead of assuming that the effect is either immediate or delayed, a different (sixth) approach consists in using Bayesian analysis to identify when is the most likely moment in which the behavior changes and, thus, whether this change is immediate or not (Natesan & Hedges, 2017; Natesan Batley, Minka, et al., 2020).

In summary, on the basis of the literature reviewed, a potential conceptual definition of immediacy could refer to a continuous feature such as the latency of the onset of the effect, regardless of whether this effect is abrupt (as a change in level) or gradual (as a change in slope), and regardless of its duration (i.e., whether it is maintained or temporary). Even if consensus could be reached regarding a conceptual definition, there are many possibilities to translate this conceptual definition into an actual quantification.

Immediacy: Operational Definitions

How Many Data Points to Include

Following the recommendation from the What Works Clearinghouse standards (Kratochwill et al., 2013), a comparison between the last three data points from the baseline phase versus the first three data points from the intervention phase has been highlighted in literature (Aydin & Tanious, 2022; Epstein et al., 2021; Gage & Lewis, 2013; Geist & Hitchcock, 2014; Haegele & Hodge, 2015; Horner & Kratochwill, 2012; Maggin et al., 2013; Michiels & Onghena, 2019; Ninci, 2019). Plavnick and Ferreri (2013) even refer to these same measurement occasions when defining how the assessment of level (rather than immediacy) should be performed. Nonetheless, there have also been calls to consider the contextual factors when determining how many data points reflect immediacy (L.-T. Chen et al., 2015).

A second option, referring to using more than three measurements per phase, is to use the last five data points from the baseline phase as compared to the first five data points from the intervention. This option was suggested by Natesan and Hedges (2017), introducing the Bayesian unknown change point analysis. It was also mentioned by Wolfe et al. (2019), when presenting a protocol for performing visual analysis, and by Barton et al. (2019) in a study comparing several data analytical options in the context of real data.

A third option, in this case referring to fewer than three measurements can also be found in literature. Specifically, Lane and Gast (2014) refer to "absolute level change" quantifying the immediacy of change from the last session of baseline to the first session during

intervention. Similarly, Ledford et al. (2018) describe an example of an immediate effect as "the first data point in each condition was different in level than data point in the preceding condition, in the expected direction" (p. 12).

As a fourth option, when discussing the point of change, Morley (2018) states that the assessment can be performed considering one, two, or three points on either side of the introduction of treatment. Thus, this agrees with both the assessment of the "absolute level change" (Lane & Gast, 2014) and with the What Works Clearinghouse (2017) recommendation.

As a fifth option, a less specific number of measurement occasions is also present in Lane and Gast (2014). These authors refer to "relative level change" quantifying the proportional change from the last half of baseline to the first half of the intervention condition, using median values.

Finally, it should also be kept in mind that Michiels and Onghena (2019) found that randomization tests for the comparison between means of complete phases are particularly sensitive to immediate effects (in random intervention point designs). Thus, when selecting between using all the data or only part of the data for quantifying immediate effects, it is important to consider two aspects. First, in case the baseline level is expected to be stable, the mean of all the baseline data could be a useful summary and there would be no need to discard baseline data. In contrast, the researchers may prefer to focus on the last baseline measurements, when any initial variability is expected to be stabilized, following typical methodological recommendations (Kazdin, 2021; Ledford et al., 2019). Second, in case the immediate effect is expected to be maintained, the mean of all the intervention phase data could be a useful summary. In contrast, if the immediate effect is expected to be decaying, it might make sense to discard later intervention phase measurements.

Overall, in case phases were very short (e.g., including only three measurement occasions), choosing between three, four, or five data points for the assessment would have been inconsequential. However, reviews of SCED data suggest that most baselines contain five or more measurements (Shadish & Sullivan, 2011), with a mean of 10 measurements in some reviews (Smith, 2012), and a median of seven in others (Pustejovsky et al., 2019). Thus, how many measurements per phase to consider when assessing immediacy is not a trivial question.

What is the Focus of the Assessment of Immediacy?

In previous literature (e.g., Kratochwill et al., 2013; Ledford et al., 2019; Maggin et al., 2018), immediacy has been considered to be one of the six visually inspected data aspects. However, it could also be considered as a "meta-aspect", in the sense that we can have an immediate effect on the level, an immediate effect on the variability, an immediate effect on the trend, and an immediate effect on the overlap. In that sense which data aspect is the focus of the assessment of immediacy is relevant for the operational definition of immediacy. Specifically, apart from level, a quantification may also take trend into account, although this is not equivalent to quantifying an immediate effect on trend. For instance, in piecewise regression, the first intervention phase measurements as predicted from the baseline trend are compared to the first intervention phase measurement as predicted from the intervention phase trend, following the logic of piecewise regression (Center et al., 1985; Moeyaert, Ugille et al., 2014). This logic is also applicable to multilevel models (Baek & Ferron, 2013). Also, in the context of regression-based analyses, it has been highlighted that change may not always be immediate, and the design matrix needs to be modified to accommodate a different expectation about the timing of the effect (Miočević et al., 2020).

A further option is to take both trend and variability into account. Specifically, in the visual aid proposed by Manolov and Vannest (2019), the immediate effect is assessed, comparing the first three intervention measurements to a projection of baseline Theil-Sen trend with a variability band (computed via the mean absolute deviation of the baseline data).

A quantification that entails extrapolating baseline trend, and combining change in level and slope, in a similar way as the generalized least squares model by Swaminathan et al. (2014), is called the "mean phase difference". For this quantification, there is an option for a limited projection of baseline trend in order

to avoid unreasonable predictions (Parker, Vannest, Davis, & Sauber, 2011), effectively leading to a quantification of an immediate effect. The exact extent of this projection, that is, the number of initial intervention phase measurement occasions considered, depends on the baseline phase length and on the degree of fit of the trend line to the baseline data, which is related to the baseline data variability (Manolov et al., 2019).

A Different Perspective on Immediacy

In the previous sections, we referred to a variety of ideas regarding how many data points to include when assessing immediacy and which data aspects to focus on. In the current section, we refer to a proposal that entails using all data and takes possible trends into consideration. Beyond the recommendations of the What Works Clearinghouse (2017), to the best of our knowledge, the only proposal dealing specifically with immediacy is Natesan and Hedges' (2017) Bayesian unknown change-point model. The result of this model is a range of moments in time in which the largest behavioral change took place with greatest probability. If this range is narrow and centered around the moment of change in phase, there is stronger evidence for an immediate effect. Alternatively, this model can help in identifying a delayed effect, if the range is posterior to the moment in which the intervention was introduced. The Bayesian unknown change-point model offers an operational definition that is apparently closer to Morley's (2018) use of the expression "point of change", which could be understood as looking at when the largest behavior change occurs and whether this change is simultaneous with the introduction of the intervention.

The main limitation of the model is its complexity, illustrated in several features. First, it is necessary to make assumptions about the outcome being continuous and normally distributed (Natesan Batley et al., 2021). Second, it is necessary to estimate autocorrelation, which can be problematic when the number of available measurements is insufficient (Huitema & McKean, 1991; Krone et al., 2016; Shadish et al., 2013). Third, it is necessary to specify priors and using inappropriate priors may lead to inappropriate results, especially in small samples (Natesan, 2019). Similarly, Rindskopf (2014) alerts that Bayesian methods "will not converge if they are not given somewhat informative priors and 'reasonable' start values." (p. 587). Finally, Bayesian methods entail data simulations using relatively complex software (Natesan Batley, Contractor, & Caldas, 2020). Even the proponents of Bayesian analysis recognize that it requires a learning process that is likely to make this proposal less attractive to applied researchers (Natesan, 2019; Natesan Batley, Contractor, & Caldas, 2020). Therefore, in the following section, we focus on an alternative that assumes a known change-point and that is distribution-free, one that is easier to understand and implement.

Proposing a Sensitivity Analysis with Randomization Test (SART) Approach

Sensitivity Analysis in Relation to the Number of Data Points (and the Focus)

From the text so far, it can be deduced that there are multiple options for specifying how many data points to use for investigating immediacy. Thus, immediacy can be assessed by comparing the results of several possible numbers of values per phase. For instance, specifying three measurements per phase as a value for assessing immediacy has been recognized as arbitrary (Kratochwill et al., 2013) and it is not the only possible option, as using one, three, or five values per phase has already been mentioned. Logically, there is no sound reason why two or four measurements per phase should not be considered, too. Thus, there are researcher degrees of freedom in the data analytic process (Wicherts et al., 2016) and it is necessary to consider their effect on the conclusions. Trying out different options and checking the degree to which the conclusions agree is well-aligned with the logic of sensitivity analysis (Steegen et al., 2016); that is, checking the degree to which several data analytical options agree or converge, although we are not specifically referring to assumptions and to robustness (e.g., Baek & Ferron, 2013; Moeyaert, Ferron, et al., 2014), and whether meeting them or not affects the validity of the conclusions.

In terms of the focal data feature, if comparing level (e.g., means or medians) is not the only possible kind of quantification, then several other options need to be considered. If trend is to be taken into account, apart from

the analyses already mentioned in the previous section, another option arises. Specifically, regression analysis can be used to detrend the baseline and intervention phase data (Allison & Gorman, 1993; Parker et al., 2006) and the residuals from each phase can be used for quantifying the immediate effect. Overall, we consider that two approaches are reasonable: (a) compute the type of quantification that represents best the kind of effect expected, following the logic of randomization tests (Heyvaert & Onghena, 2014b; Levin et al., 2021), and/or (b) follow a sensitivity analysis approach, computing several possible quantifications and checking the degree to which their conclusions converge. Whenever there is a theoretical or empirical basis for any expectations about the data pattern, we advocate for deciding a priori the focal data feature and following the sensitivity analysis approach only for the number of measurements involved. However, in case the prespecified choice of a test statistic is inappropriate (e.g., a mean difference was selected before gathering the data, but clear trends appear in the data once collected), then it would be reasonable to also report the results using another test statistic, selected a posteriori. This would be an example of using visual analysis to validate the results of a statistical analysis (Parker et al., 2006) and it would require making explicit the fact that the choice of the (second) test statistic is made a posteriori and how the results using different test statistics differ.

Randomization Test Logic

Assessment of Effects in the Presence of Randomization

Randomization in the design and the use of randomization tests for analyzing the data were initially suggested by Edgington (1967, 1996). Afterwards, the strengths of randomization tests have been echoed by multiple authors (Craig & Fisher, 2019; Jacobs, 2019; Kratochwill & Levin, 2010; Onghena, 1992). The idea of the proposal is to benefit from the use of randomization in the design. Specifically, the type of randomization refers to selecting at random when to change the phase, with certain restrictions (i.e., minimum phase length of 5 as per the WWC standards), as described by Edgington (1975). On the basis of the randomization in the design, all possible divisions of the data (i.e., moments of phase change) could have taken place and all are equally likely under the null hypothesis of no intervention effect. The question that can be answered via the randomization test is whether a difference as large as or larger than the one actually observed is exceptional, given that the null hypothesis is true. Thus, the descriptive result (e.g., a mean difference) is compared to other possible results (e.g., mean difference for all the admissible points of change in phase) in order to quantify the degree to which the former is expected to happen in absence of an effect. If the probability is small (e.g., equal to or smaller than .05), then there is evidence of the statistical significance of the result.

Caution in the Absence of Randomization

Following the logic of randomization tests, it is possible to suggest a permutation test (in the absence of actual random assignment) in order to perform calculations analogous to the ones performed when a randomization test is used. Thus, an exploratory assessment of the degree to which there is evidence about an immediate effect is also possible in the absence of randomization (Edgington & Onghena, 2007; Onghena et al., 2019). However, the statistical-conclusion validity relies on randomization actually taking place in the design. Therefore, any interpretation of the *p*-value resulting from the permutation test should be made with caution. Nevertheless, this limitation is also applicable when using other kinds of statistical tests that rely on random sampling which has not actually taken place.

Test Statistic

When the phases are compared, the difference is quantified via a test statistic. In terms of the data aspect that is the focus of the quantification, this test statistic can be a difference in means, a difference in slopes, or a nonoverlap index (Heyvaert & Onghena, 2014a). Ideally, the choice of what data aspect to quantify should be based on the expected data pattern: mean difference when stable data are expected, change in slope when a progressive linear effect is expected, and a nonoverlap index in case the measurement of the target variable is expressed in ordinal terms and an

interval or ratio scale cannot be assumed. In terms of the number of data points to include when computing the test statistic, it is possible to use all measurements or only some of them (e.g., the last three of the baseline phase and the first three of the intervention phase). If there is no clear justification for choosing a given number of measurements, a possibility is to include several possible values and check the extent to which the results differ.

Numerical Summary

The value computed for the actual point of change in phase is called "the observed value of the test statistic", whereas the values computed for all other admissible points of change in phase are called "pseudovalues" or "potential values". A way to assess the relative size of the observed value is by ranking all pseudovalues and the observed value itself in ascending order (from smallest to largest). From the rank of the observed value, the randomization *p*-value can easily be determined as the proportion of pseudovalues as large as or larger than the observed value of the test statistic. If desired, it is possible to determine whether the observed value of the test statistic is one of the 5% largest (if there are at least 20 admissible intervention start points), or whether it is the largest one.

Combining the Sensitivity Analysis with Randomization Test Logic

The randomization test logic can be followed for checking the degree to which the quantification of the difference between conditions in the focal data feature (e.g., a mean difference) is among the largest ones that could have been obtained for all admissible points of change in phase. Complementarily, the sensitivity analysis approach can be followed for checking the degree to which the extremeness of this quantification is similar across different operational definitions of the immediate effect (i.e., different number of measurements used). Thus, if the quantification for the actual point of change in phase is among the (5%) largest ones and this quantification is similarly among the largest ones regardless of the number of measurements used for computing, then the evidence for an immediate intervention effect would be stronger. In summary, the sensitivity analysis with randomization test (SART) approach can be conceptualized as a tool for identifying when changes were most clear and whether the moment of maximal change warrants a conclusion of an immediate effect.

Graphical Representation

The numerical results can be accompanied by a graphical representation, similar to a histogram. In this chart, each admissible point for change in phase, taking into account possible restrictions for a minimal phase length (Edgington, 1996) would be represented on the abscissa (X-axis). On the ordinate (Y-axis), the observed value of the statistic and the pseudovalues are represented, including a horizontal line for 0, representing no difference between the phases. Positive values (above the horizontal line) are "favorable" results, that is, results in line with the expected effect or the alternative hypothesis. Furthermore, the observed value of the test statistic is depicted in green if a favorable result is obtained. It is depicted in red if an unfavorable result is obtained.

Correspondence with Conceptual Definitions

Apart from obtaining the rank or the *p*-value for the statistic, thanks to the proposed graphical representation, it is possible to observe the distribution of the differences and check around which potential moment of change the differences are larger. Moreover, it is possible to identify the specific moment for which the change is largest. Both these pieces of information allow assessing whether the largest change occurs when the intervention is introduced or the effect is delayed, and by how many measurement occasions. Therefore, it is possible to evaluate the quickness of change, which is well-aligned with most conceptual definitions of immediacy. In contrast, merely comparing the last three (or in general, k) baseline measurements to the first three (k) intervention phase measurements, only quantifies an abrupt change in level but does not inform about quickness.

Illustrations with A-B designs

At this stage, we will present a series of examples of the SART approach. For the sake of simplicity, in all examples, we will focus on the mean level, rather than on trend, variability, or overlap. However, in real research, the appropriate practice would be to choose a test

statistic (and a focal data feature) prior to gathering the data, on the basis of the expected effect of the intervention (Heyvaert & Onghena, 2014b; Levin et al., 2017), according to the theoretical background and the empirical evidence available. Additionally, in all examples we will assume that the minimum number of measurements per phase is set at five, following common recommendations (Tate et al., 2013; What Works Clearinghouse, 2020).

The first illustrative data set is an AB design with randomly chosen intervention start point (Winkens et al., 2014). The aim of the study was to reduce the number of aggressive behaviors in a person diagnosed with olivo-ponto-cerebellar ataxia, using a method that entails identifying antecedent events, target behaviors, and consequent events. Figure 1 represents the raw data, whereas Figure 2 represents the output of the SART approach.

The visual inspection of Figure 1 suggests that there is a clear immediate effect if the last three baseline measurements are compared to the first three intervention phase data points. However, this effect is temporary, as there is a deterioration in later sessions and also increased variability.

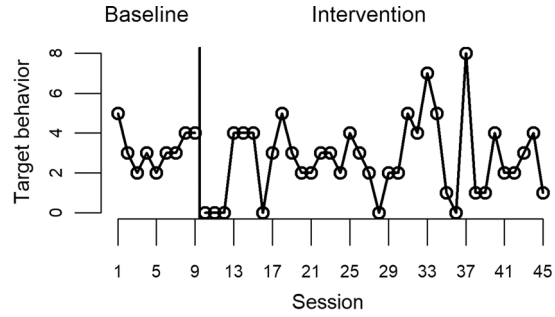
Figure 2 offers a representation of how the rank of the value of the test statistic varies according to the number of data points included from each phase. Specifically, if three measurements per phase are included, the value of the test statistic is the largest of all possible data divisions. In case more than three or less than three measurements per phase are used, the results are fairly consistent in that the value of the test statistic is one of the largest (although not the largest). When more measurements are included, there is less evidence for immediate effect, which agrees with the inspection of the time series line plot: The effect is lost with time. Finally, it should be noted that beyond the immediate effect, the complete time series resembles noise (both looking at the time series plot from Fig. 1 and the bottom right plot from Fig. 2).

The following three datasets were used by Natesan Batley et al. (2021) illustrating the Bayesian Unknown Change Point analysis for three different data patterns. In their article, these data were named Dataset 1 (Coulter & Lambert, 2015), Dataset 2 (Macpherson et al., 2015), and Dataset 3 (Barber et al., 2016).

The data gathered by Coulter and Lambert (2015) are the percent of correct words read

Figure 1

Winkens et al. (2014) Data

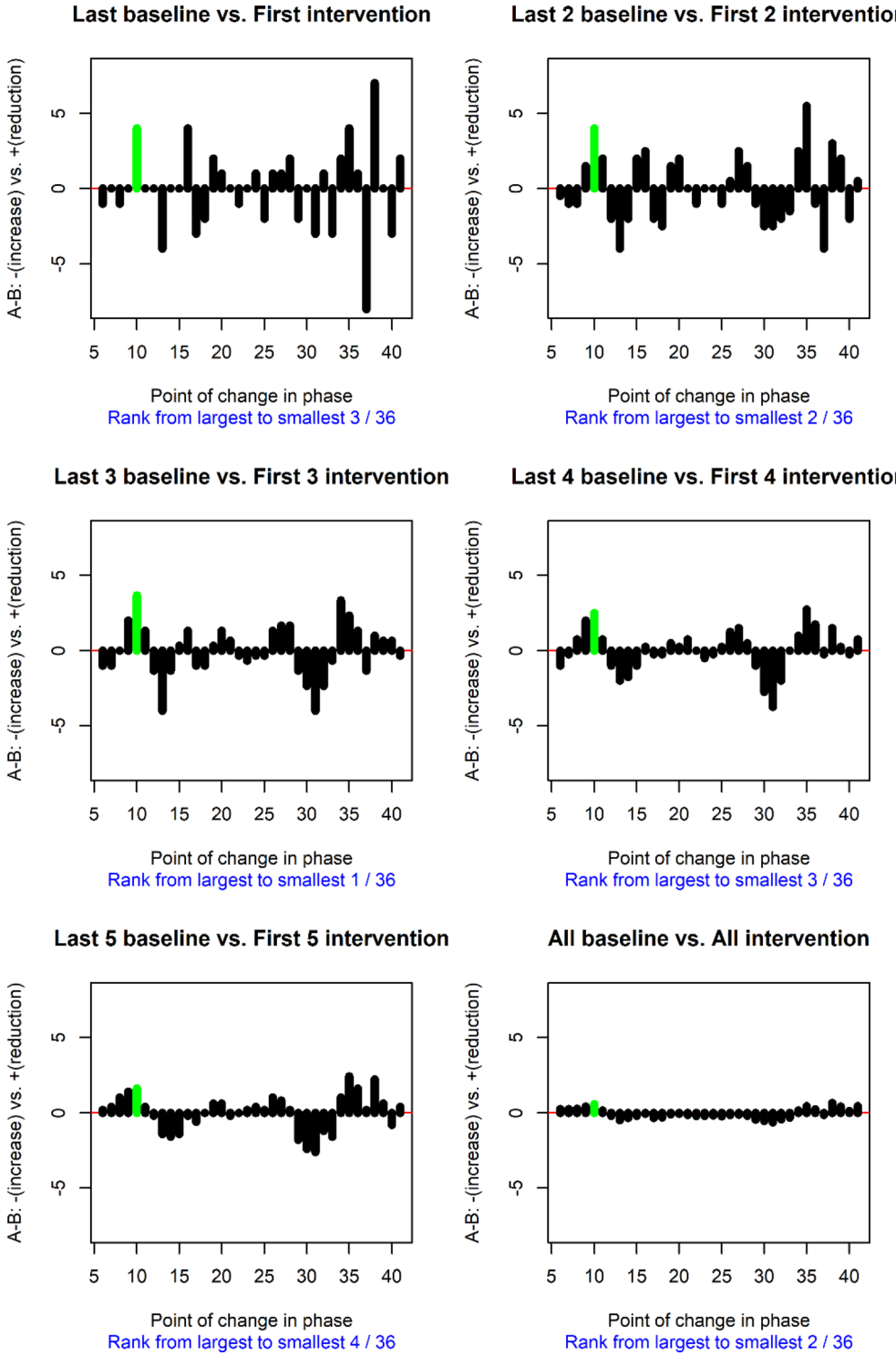


per minute by participants with a learning disability. The aim was to increase this target behavior, using preteaching of keywords. The raw data are depicted in Figure 3, whereas Figure 4 represents the output of the SART approach. The data from Figure 3 suggest an immediate effect, occurring simultaneously with the change in phase at measurement occasion 13. When inspecting Figure 4 we see that the evidence for an immediate effect is stronger (i.e., the value of the test statistic is among the largest one or two) when three or fewer values per phase are used to compute the immediate effect. When more values are used, the immediate effect does not look as evident. Moreover, all panels of Figure 4 agree that there is an additional improvement at the end of the data series, after measurement occasion 26.

The data gathered by Macpherson et al. (2015) are the percentage of opportunities in which verbal compliments occurred, as expressed by a child diagnosed with autism. The aim was an increase in this target behavior, using portable video modeling technology. The raw data are depicted in Figure 5, whereas Figure 6 represents the output of the SART approach. The visual inspection of Figure 5 suggests that the latency of the effect is two measurement occasions (i.e., it is not immediate). Moreover, this delayed effect is not abrupt, but rather somewhat gradual in that the upper asymptote is not reached initially at the third intervention phase measurement occasion, but rather at the fourth one. Accordingly, the SART approach panels in Figure 6 suggest that the mean difference for the actual moment of change in phase is not

Figure 2

Results of the Randomized Sensitivity Analysis, Using a Mean Difference as a Test Statistic, for the Winkens et al. (2014) Data



among the largest ones; the pseudovalues of the test statistic are larger for later measurement occasions.

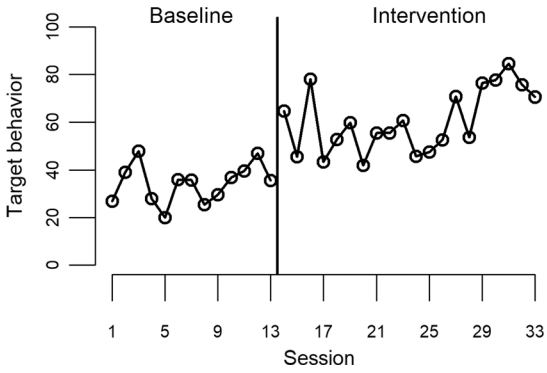
The data gathered by Barber et al. (2016) represent social communication skills of a preschool child diagnosed with autism. The aim was to increase the number of responses, using peer mediated interventions. The raw data are depicted in Figure 7, whereas Figure 8 represents the output of the SART approach. The visual inspection of Figure 7 suggests that there is a gradual increase of the target behavior throughout the intervention phase, until measurement occasion 17 when an apparent upper asymptote is reached. However, this gradual improvement appears to be the result of an already existing improving baseline trend. The SART approach panels in Figure 8 suggest that the mean difference for the actual point of change in phase is not among the largest ones, further emphasizing the lack of an immediate effect. Actually, the largest changes are observed if the point of change in phase were earlier (measurement occasion 6) or later (measurement occasions 16 or 17). Finally, when all data are taken into account (bottom right panel of Fig. 8), it is apparent that the mean difference remains practically unchanged for all admissible points of change in phase, suggesting that there is no clear effect of the intervention.

Illustration with Methodologically Appropriate Single-Case Experimental Designs

Description of the Data. For the current illustration, we use the data from a multiple-baseline design across participants, as reported by Laski et al. (1988). Eight children diagnosed with autism spectrum disorder participated, four being nonverbal (see Fig. 9) and four presenting echolalia (see Fig. 10). The aim of the study was to increase the children's speech by using the Natural Language Paradigm, and the target behavior is measured as the percentage of intervals with child vocalizations. There are two data sets for one of the nonverbal children: one interacting with the mother and one with the father. We also use the data from an ABAB design, replicated across participants, as reported by Lambert et al. (2006). The disruptive behavior (quantified as the number of intervals present) of four fourth-grade students was studied, with Condition A representing single-student responding and Condition B being the response card treatment. In the Lambert et al. study, there are four participants in Class A and five participants in Class B, but for the current illustration we included only the data from two participants from Class B (see Fig. 11). Both data sets have been used in previous illustrations of data analytical approaches: Lambert et al. in a *Journal of School Psychology* special issue (Shadish, 2014), and also in Peng and Chen (2018); Laski et al. (1988) in Hedges et al. (2013), Moeyaert et al. (2014), and Natesan and Hedges (2017).

Figure 3

Coulter and Lambert (2015) Data



Analyses Performed. The analyses consisted in computing mean differences for each admissible data division, considering a minimum of three measurement occasions per phase. The rank of the mean difference for the actual moment of change in phase was obtained with lower ranks indicating greater difference (i.e., the rank of 1 corresponded to the situation in which the mean difference was largest for the actual moment of change in phase). We also noted for which potential moment of change in phase the mean difference was largest: it could be the actual one, or a previous or later moment.

A sensitivity analysis was performed, computing these mean differences, as well as the ranks and the moments of largest difference, for different amounts of data. Specifically, we compared the results of using one, two, or three measurements per phase (referring to the last baseline and the first intervention phase data points), and using for all the measurements available.

Figure 4

Results of the Randomized Sensitivity Analysis, Using a Mean Difference as a Test Statistic, for the Coulter and Lambert (2015) Data

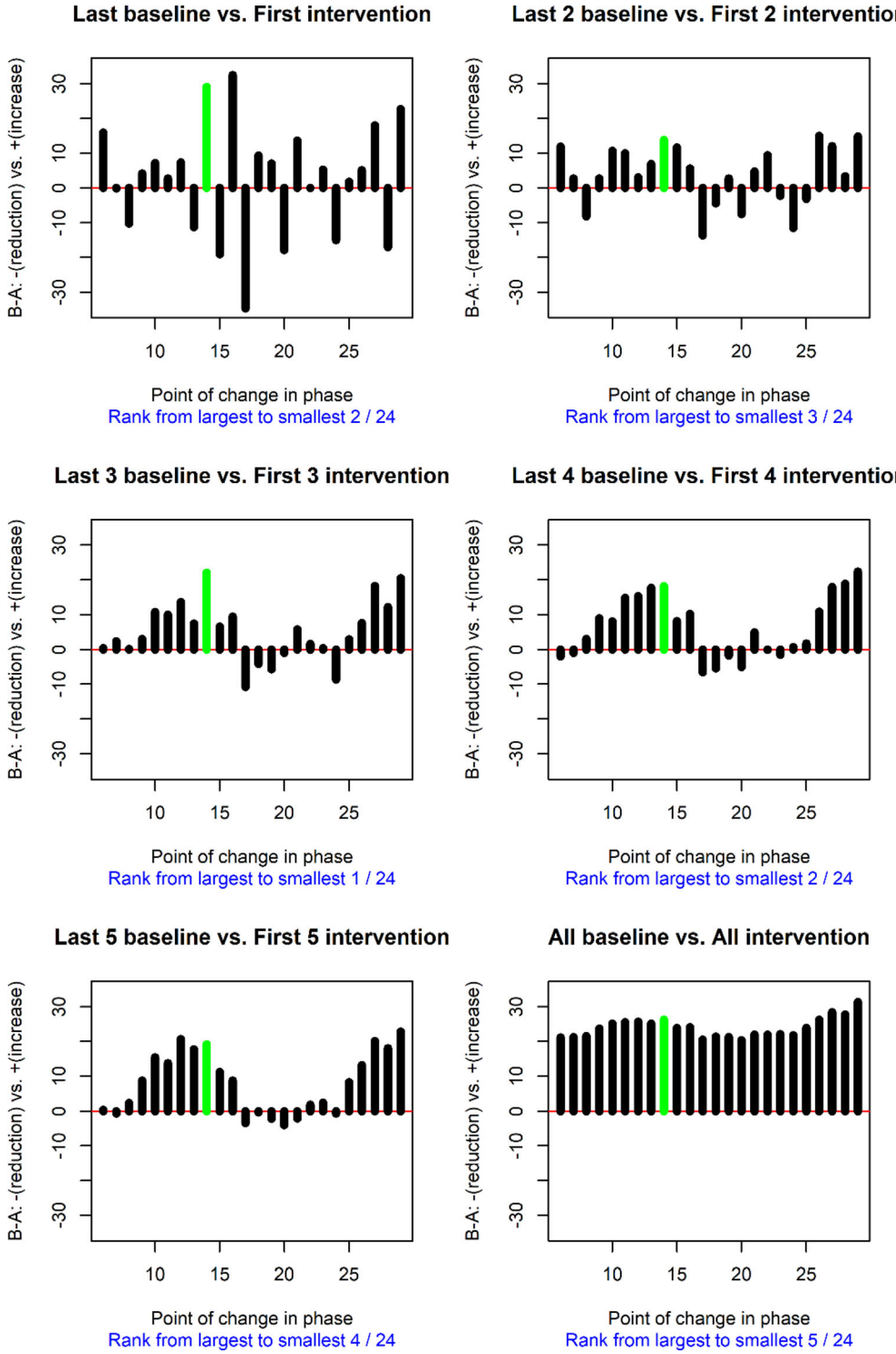
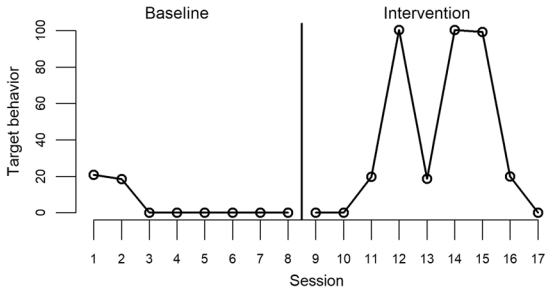


Figure 5

Macpherson et al. (2015) Data



For the Lambert et al. (2006) data, we performed three comparisons per participant: A_1-B_1, B_1-A_2, and A_2-B_2. For the Laski et al. (1988) data, there was one A-B comparison per tier.

In the subsequent section, we represent a numerical summary of the results, constructed on the basis of graphical representations similar to Figures 2, 4, 6, and 8.

Results for the Laski et al. (1988) Data. Table 1 includes the results for the Laski et al. (1988) data. The cells are colored in order to aid the readability. Green cells mark participants for whom the largest difference was observed for the actual moment of change in phase (i.e., nonverbal Participants 2 and 4; Participants 1, 2, and 4 with echolalia). For these participants, the evidence for an immediate effect is strongest, especially considering that the number of measurements used for computing the mean difference does not affect the result.

In red are marked the cells that represent participants for whom the largest mean difference is for a potential moment of change in phase that is later than the actual moment. This is the case for nonverbal Participant 3, for whom the mean difference for the actual moment of change in phase is among the largest ones. In contrast, for Participant 3 with echolalia, according to the amount of data used for computing the mean difference, the largest difference is observed for either earlier (cells in orange) or later (cells in red) potential moments of change in phase. In any case, the mean difference for the actual moment of change in phase is among the smallest ones.

Finally, for nonverbal Participant 1, the largest mean difference is observed for earlier potential moments of change in phase (cells in orange), and the actual mean difference is among the second largest.

Overall, there is evidence for a consistent immediate effect for three of the four children with echolalia. For the nonverbal children, the mean difference for the actual moment of change in phase is the largest (or one of the two largest) for all participants, suggesting that the effect could be considered immediate.

Results for the Lambert et al. (2006) Data. Table 2 contains the results for Participants B3 and B4 from the Lambert et al. dataset. The green cells for Participant B3 suggest that there was an immediate reduction of the target behavior with the introduction of the intervention (i.e., comparisons A_1-B_1 and A_2-B_2), but there was not an immediate increase of the target behavior with the withdrawal of the intervention (i.e., comparison B_1-A_2). Actually, the largest increase is taking place in later measurement sessions (cells in red) in phase A_2. This could be understood as evidence for a delayed or progressive deterioration, once the intervention is withdrawn.

For Participant B4, there is no evidence for an immediate effect, as the largest mean difference is observed always for potential moments of change in phase, which are earlier than the actual moment. Moreover, the mean difference for the actual moment of change in phase is not among the largest ones. Thus, any changes in the behavior cannot be attributed to the change in conditions.

Overall, if we consider only the results for these two participants (and not for all nine participants, whose results are reported by Lambert et al., 2006), the evidence for an immediate effect is not consistent: partially present for one of the participants (although deterioration is rather gradual) and not present for the other participant.

Alternative Analyses. It should be noted that the illustrations presented so far used a mean difference as a quantification, or test statistic in a randomization test (if we convert the ranks to *p*-values and in case the moment of change in phase was actually determined at random). This should not be interpreted as a recommendation of using only mean difference as a quantification. We rather echo the recommendation to choose the test statistic according to the type of effect expected (Edgington, 1975), for instance a difference in

Figure 6

Results of the Randomized Sensitivity Analysis, Using a Mean Difference as a Test Statistic, for the Macpherson et al. (2015) Data

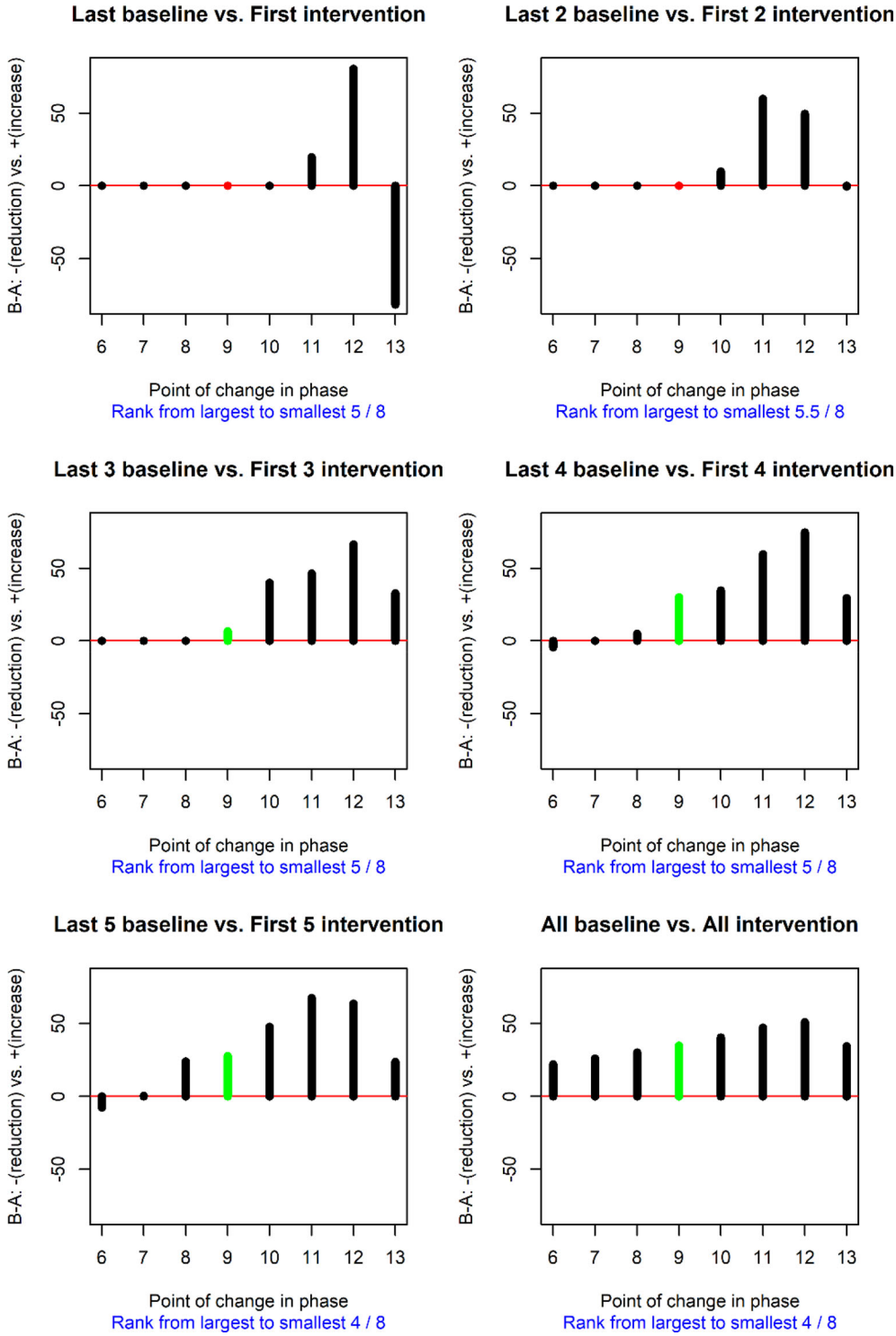
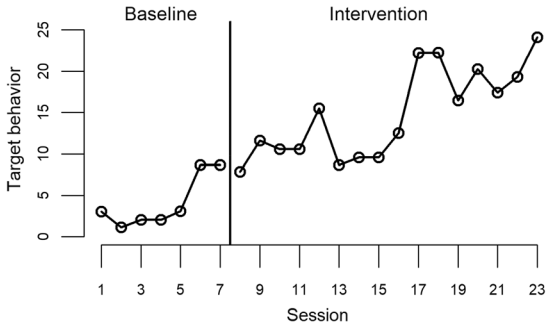


Figure 7

Barber et al. (2016) Data



slope or variability (Levin et al., 2021), or a delayed effect (Levin et al., 2017). Thus, the choice of a quantification is based on the subject-matter knowledge of the kind of target behavior and the intervention used. In relation to immediacy, if we consider it a "meta-aspect", there could be an immediate change in level or an immediate change in slope, or an immediate nonoverlap (if the researcher is using ordinal data or is not willing to summarize / represent the data via flat mean lines or trend lines).

Peculiarities Arising from Using Complex Designs. When applying the SART approach to designs that are more complex (and more appropriate than AB), there are two aspects to be taken into account. On the one hand, a multiple-baseline or a reversal design can be decomposed into a series of A-B comparisons, each of which is first analyzed separately and afterwards the consistency across all within-study replications (i.e., all A-B comparisons within the multiple-baseline or a reversal design) is assessed. This logic is consistent with the idea that each A-B comparison can be considered a "basic effect" (Horner & Odom, 2014), which is necessary but not sufficient, as the consistency of effects is to be evaluated (Kratochwill et al., 2013). Relatedly, the assessment of consistency of effect has each A-B comparison as a building block (Tanious et al., 2020). Similarly, previous recommendations about how to pool several A-B comparisons when assessing intervention effectiveness have referred to counting the number of positive results in each A-B comparison and assessing whether a 3:1 ratio of effects to no effect is achieved (Cook et al., 2015; Maggin et al., 2013).

On the other hand, to better understand the complete data pattern, it is necessary to consider the possibility that all transitions between phases are not exactly equivalent in a reversal design or that a within-subject replication from a reversal design may not be the same as a between-subject replication in a multiple-baseline design. Regarding the reversal designs, it is possible to have an a priori expectation for an extinction burst (Katz & Lattal, 2021) during the first A-B comparison in an ABAB design and the possibility for an immediate effect in the second A-B comparison. Regarding multiple-baseline designs, between-cases (across-series) comparisons may be relevant for verification (Carr, 2005; Ferron et al., 2014), which would entail that it is not sufficient to assess each A-B comparison separately. However, it has been recently emphasized that between-cases comparison, typical for concurrent multiple-baseline designs, is not a critical aspect (Ledford, 2022; Slocum et al., 2022).

In order to address and accommodate complexities within the SART approach, two aspects are crucial: (a) the explicitly stated a priori expectation (with its justification) about the kind of effect of introducing or withdrawing an intervention for each participant; and (b) the evidence obtained regarding the moment in which the largest change is observed. The a priori expectation and the empirical evidence are then compared in order to assess the degree to which there is support for a causal inference.

Boundary Conditions When Using Complex Designs. The SART approach is applicable to designs that consist of phases (i.e., several consecutive measurements in the same condition), in order to be able to assess when the largest change is taking place. In that sense, in an alternating treatment design (which is expected to be used when target behavior is susceptible to immediate changes, coinciding with the rapid change in conditions), the application of SART would not be possible, as the common recommendation is for a maximum of two consecutive measurements per condition (Kratochwill et al., 2013). A changing criterion design does include several measurements per condition, but achieving immediately the criterion level desired may

Figure 8

Results of the Randomized Sensitivity Analysis, Using a Mean Difference as a Test Statistic, for the Barber et al. (2016) Data

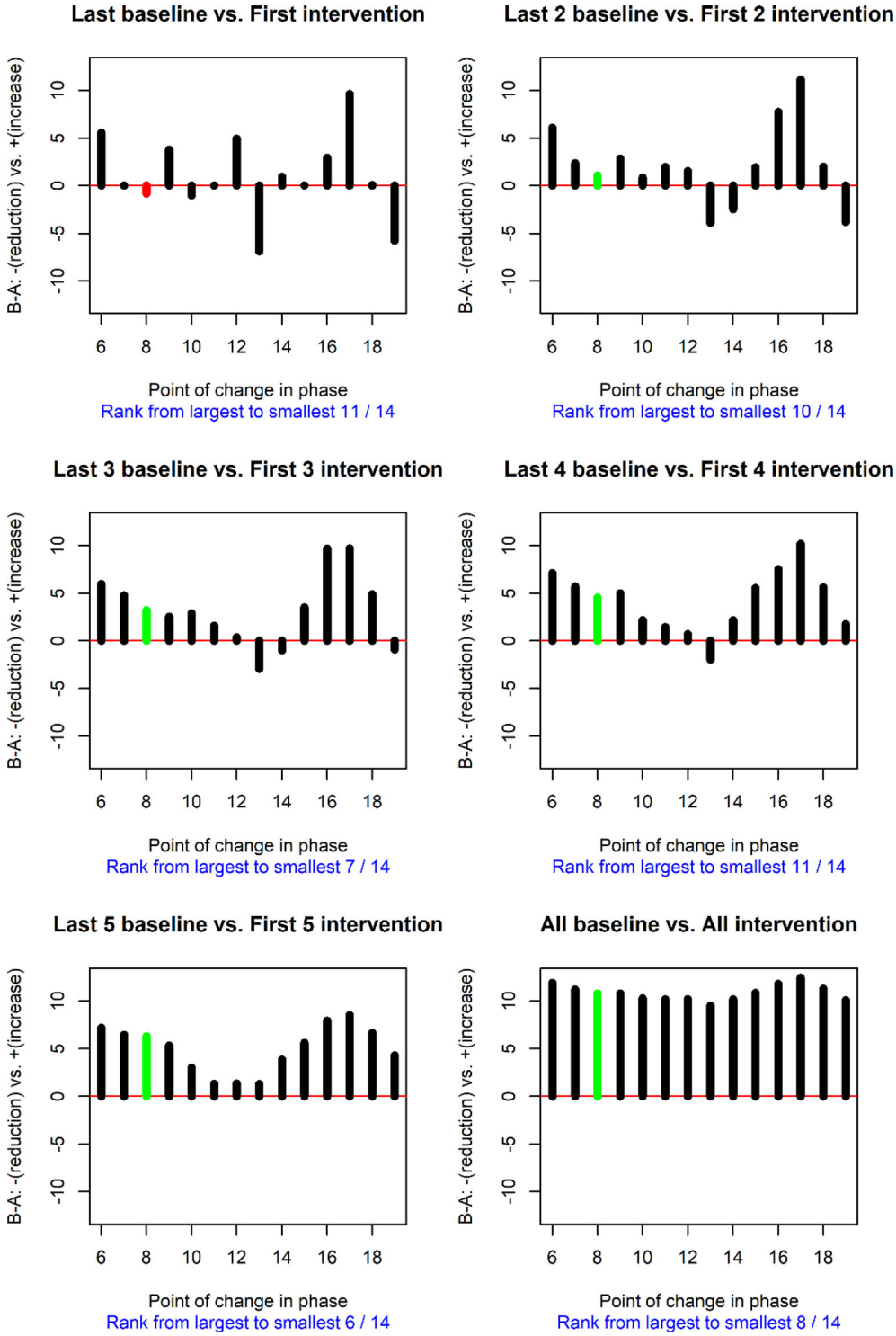
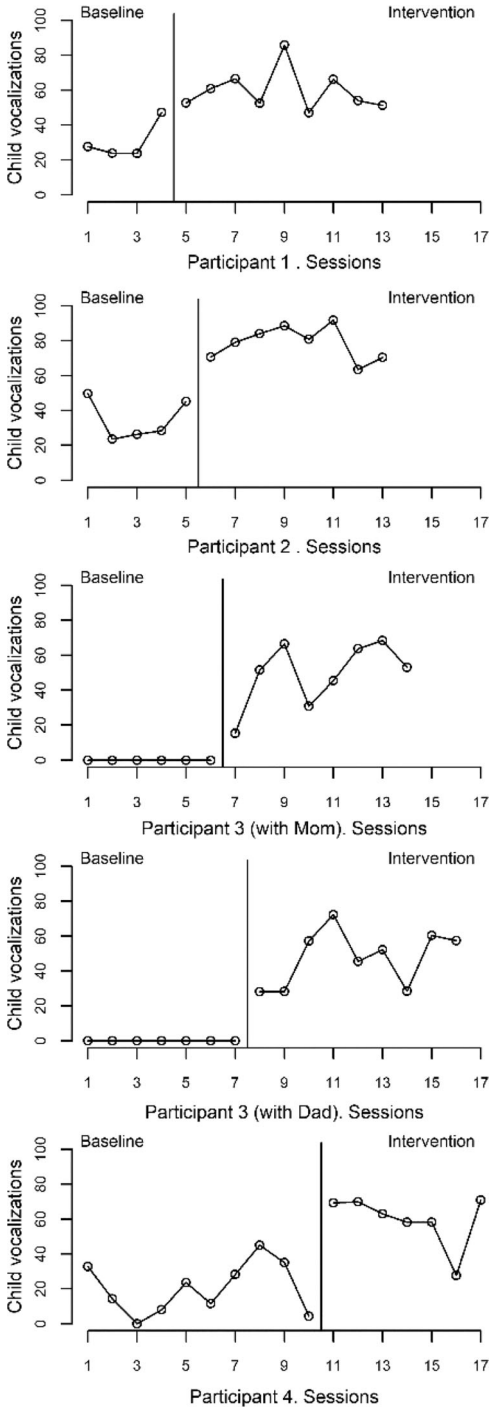


Figure 9

Multiple-Baseline Data for Nonverbal Children, Gathered by Laski et al. (1988)



not be critical in case the range-bound version (McDougall et al., 2006) is used, specifying an acceptable range around this criterion level. Moreover, the aim may not be to achieve the pre-established level immediately, but to eventually achieve a mastery criterion, before changing the criterion level (Manolov et al., 2020).

Discussion

Assessing Immediacy and Inferring Causality

Immediate Effects

In order to be able to state that the observed data pattern is indicative of a causal effect of the intervention, the explicitly stated a priori expectations of the researchers are crucial because in that way observations are firmly linked to a causal theory or a body of previous research findings. The researcher needs to state whether an effect is expected to be immediate or to have a certain delay/latency. Such an expectation needs to be justified on the basis of previous research or theory. The assessment of the potential moment of change in phase for which the difference is largest (i.e., the rank is equal to 1) can be understood as another way of checking whether the actual moment of change in phase coincides with the largest difference, suggesting an immediate effect.

Delayed Effects

It is relevant to consider the possibility of transition states, involving a change from one stable state of behavior to another (Sidman, 1960). A transition state is expected to start when a change in the conditions is introduced, but this also depends on: (a) the kind of intervention (e.g., aiming extinction of a behavior by means of no longer rewarding it; fixed- vs. variable-interval reinforcement schedule), which affects the perception of the participant regarding when conditions have changed; and (b) the moment in which a just-noticeable effect is produced. A specific kind of transition state, called "extinction burst" is especially relevant when referring to immediacy. An extinction burst entails an immediate effect that is the opposite of the desired one, in relation to the use of extinction as intervention to reduce an undesirable behavior (Katz & Lattal, 2021).

Figure 10

Multiple-Baseline Data for Children with Echolalia, Gathered by Laski et al. (1988)

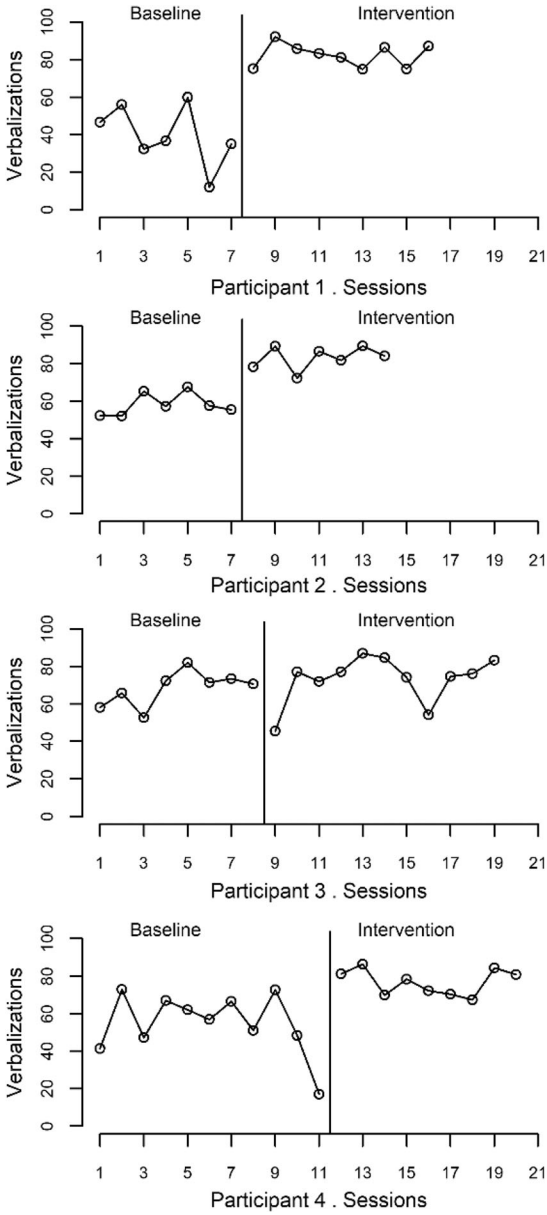
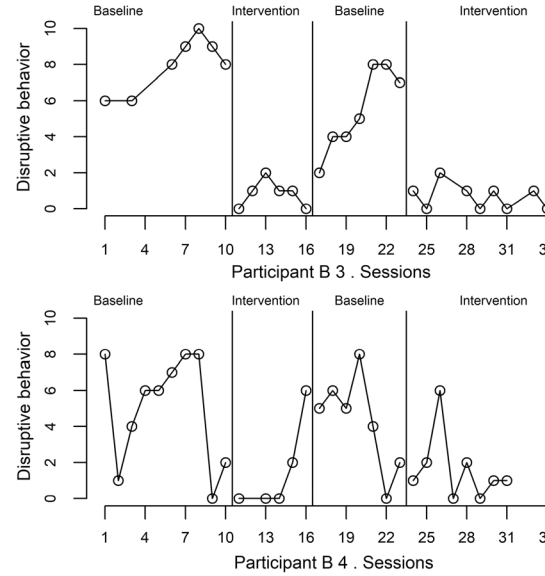


Figure 11

Replicated ABAB Data (Participants B3 and B4), Gathered by Lambert et al. (2006)



Lack of clarity about the definition of the beginning and the end of a transition state may lead to (unnecessarily) discarding data, because of not knowing whether it belongs to a transition state or a stable state (Sidman, 1960). The approach we propose does not entail discarding data in the beginning of the new condition, due to unclear assessment of such data (once gathered). In contrast, our approach is based on the researcher's expectations regarding whether a transition state will be present or not. If a transition state is not expected for the specific intervention and target behavior (on the basis of previous research), then an immediate effect may be of interest. If a transition state is expected, then it may be more relevant to study the latency of the effect (i.e., for which potential moment of change in phase is the difference largest).

In relation to delayed effects, it would be important to assess whether a similar amount of delay is observed for all A-B comparisons (i.e., consistency of effects, in a replication within a case as in an ABAB design or across cases as in a multiple-baseline design) and whether this delay is reasonable and does not preclude inferring that the behavioral change is due to the intervention, rather than to some extraneous factor.

Finally, apart from studying delayed effects, it is also possible to obtain evidence that the behavior change is not related to the intervention, because the largest change takes places before the introduction of the intervention.

Addressing Uncertainties

The proposal addresses two kinds of uncertainty. The first kind of uncertainty results

Table 1

Results for the Multiple-Baseline Designs by Laski et al. (1988) with Five Nonverbal Children and Four Children with Echolalia

Laski - Nonverbal	P1	P2	P3Mom	P3Dad	P4
Number of possible moments of change in phase	8	8	9	11	12
Moment of change - Session #	**5**	**6**	**7**	**8**	**11**
Moment of largest difference (1 datum per phase)	**9**	**6**	**8**	**10**	**11**
Rank of observed difference (1 datum per phase)	6	1	3	2	1
Moment of largest difference (2 data per phase)	**4**	**6**	**8**	**10**	**11**
Rank of observed difference (2 data per phase)	2	1	2	3	1
Moment of largest difference (3 data per phase)	**4**	**6**	**8**	**9**	**11**
Rank of observed difference (3 data per phase)	2	1	2	3	1
Moment of largest difference (all data)	**4**	**6**	**8**	**8**	**11**
Rank of observed difference (all data)	2	1	2	1	1

Laski - Echolalia	P1	P2	P3	P4
Number of possible moments of change in phase	11	9	14	15
Moment of change - Session #	**8**	**8**	**9**	**12**
Moment of largest difference (1 datum per phase)	**8**	**8**	**10**	**12**
Rank of observed difference (1 datum per phase)	1	1	14	1
Moment of largest difference (2 data per phase)	**8**	**8**	**4**	**12**
Rank of observed difference (2 data per phase)	1	1	11	1
Moment of largest difference (3 data per phase)	**8**	**8**	**12**	**12**
Rank of observed difference (3 data per phase)	1	1	9	1
Moment of largest difference (all data)	**8**	**8**	**4**	**12**
Rank of observed difference (all data)	1	1	11	1

Note. Cells in green mark comparisons for which the largest difference is observed for the actual moment of change in phase. Cells in orange mark comparisons for which the largest difference is observed before the actual moment of change in phase. Cells in red mark comparisons for which the largest difference is observed after the actual moment of change in phase. Numbers in bold correspond to moments of change in phase.

from the definitions of immediacy and the different number of data points per phase that have to be included in the calculations. Similarly, when Sidman (1960) refers to the number of measurement occasions that need to be considered when assessing a change (including the beginning and end of a transition state) in the target behavior, he emphasizes the importance of both the number of repeated measurements available and their spacing, for making such an assessment possible. Considering that no clear-cut answer is available, the sensitivity analysis we propose (in relation to the amount of data used when quantifying differences) seems justified. A sensitivity analysis can be performed to assess the robustness of the immediate effect under different situations (i.e., different number of data points used for the quantification).

The second kind of uncertainty refers to delayed effects. In case there is not enough empirical evidence to have an expectation about the amount of delay, the approach proposed can show for which delay the evidence is strongest.

Overall, in relation to the difficulty to mark with precision when a transition ends and stability begins, Sidman (1960) highlights that "we cannot escape from the continuous temporal properties of a behavioral state by arbitrarily selecting discrete observation points." (p. 287). The study of latency (i.e., for which potential moment of change in phase the difference is largest) is well-aligned with this idea.

Relation to Visual Analysis

In terms of SCED data analysis, the assessment of the intervention effectiveness requires evaluating the adequacy of the design, the evaluation of the data gathered (Brossart et al., 2014) and social validity (Horner et al., 2005). The evaluation of the data is usually both visual and quantitative (Tanious & Onghena, 2021).

Regarding visual analysis, there are three links with the proposal. First, immediacy is one of the aspects that is supposed to be inspected visually (Kratochwill et al., 2013; Ledford et al., 2019; Maggin et al., 2018) and thus the topic is relevant for visual analysis. Second, the main guidance provided so far for the

Table 2

Results for Two of the Participants from Class B, from the Study by Lambert et al. (2006), Each Following an ABAB Design

Lambert - Participant B3	A_1-B_1	B_1-A_2	A_2-B_2
Number of possible moments of change in phase	8	8	11
Moment of change - Session #	**8**	**8**	**8**
Moment of largest difference (1 datum per phase)	**8**	**11**	**8**
Rank of observed difference (1 datum per phase)	1	2,5	1
Moment of largest difference (2 data per phase)	**8**	**11**	**8**
Rank of observed difference (2 data per phase)	1	2	1
Moment of largest difference (3 data per phase)	**8**	**9**	**8**
Rank of observed difference (3 data per phase)	1	4	1
Moment of largest difference (all data)	**8**	**11**	**8**
Rank of observed difference (all data)	1	4	1
Lambert - Participant B4	**A_1-B_1**	**B_1-A_2**	**A_2-B_2**
Number of possible moments of change in phase	10	7	10
Moment of change - Session #	**11**	**8**	**8**
Moment of largest difference (1 datum per phase)	**9**	**5**	**11**
Rank of observed difference (1 datum per phase)	2	5,5	5
Moment of largest difference (2 data per phase)	**9**	**5**	**6**
Rank of observed difference (2 data per phase)	4,5	4	7,5
Moment of largest difference (3 data per phase)	**9**	**5**	**6**
Rank of observed difference (3 data per phase)	4	5	9
Moment of largest difference (all data)	**9**	**4**	**5**
Rank of observed difference (all data)	4	5	7

Note. Cells in green mark comparisons for which the largest difference is observed for the actual moment of change in phase. Cells in orange mark comparisons for which the largest difference is observed before the actual moment of change in phase. Cells in red mark comparisons for which the largest difference is observed after the actual moment of change in phase. Numbers in bold correspond to moments of change in phase.

assessment of immediacy refers to comparing the last three baseline measurement occasions to the first three intervention phase measurement occasions (Horner & Kratochwill, 2012). However, there has been no specific visual aid related to this aspect, as there have been for level, trend, or variability. We are providing a visual representation (and a quantification) of one possible way of assessing immediacy. Third, the visual inspection of the traditional time-series line graph can be used to assess the degree to which the researcher's expectation about the type of effect (e.g., change in level) is reasonable considering the data actually obtained. Thus, visual inspection can be used as a means for validating the quantitative approach followed for summarizing numerically the magnitude of the intervention effect (Parker et al., 2006).

Repeated Arbitrary Values and Sensitivity Analysis

The typical number of measurements that have to be included when assessing immediate effects, as commonly mentioned in the methodological single-case literature, are 3 to 5 (What Works Clearinghouse, 2017; Wolfe et al., 2019). Incidentally, the values of 3 and 5 are the same as the ones for the required minimal phase lengths in past recommendations and more recent ones (Tate et al., 2013). These values are arbitrary and are likely to be overridden by the stability (or lack thereof) in the data pattern (Barnard-Brak et al., 2021) or by ethical considerations (Lancioni et al., 2021).

Moreover, the value of 3 is also the same as the number of replications required for phase designs, whereas 5 is the number of replications required for alternation designs (Kratochwill et al., 2013), but these values are also arbitrary (Wolery, 2013). Moreover, there have been some studies discussing the need for that many replications (Lanovaz & Turgeon, 2020; Lanovaz et al., 2019).

Finally, 5 is also the number of SCED studies and 3 is the number of independent research teams required for establishing an intervention as an evidence-based practice (Kratochwill et al., 2013). These values are also "somewhat arbitrary but they are based on

both expert judgment and logic" (Hitchcock et al., 2015, p. 466).

In summary, arbitrary values based on expert judgment may be useful, but they need not be considered fixed and immutable. In that sense, these values are open to discussion and proposals such as the present one aim to show the effects of different possible values on the conclusions regarding the presence of an immediate effect. Thus, a kind of sensitivity analysis is performed exploring the extent to which the evidence of an immediate effect is sensitive to data analytical choices (Steegen et al., 2016), or researcher degrees of freedom (Wicherts et al., 2016).

Implications

Number Crunching and Making Sense of the Numerical Values

Computing multiple values for multiple admissible points of change in phase and multiple possible definitions of an "immediate effect" may look like simple number crunching and lead potentially to overfitting.[1] However, checking the results for all admissible points of change in phase makes sense when randomization is used in the design and under the null hypothesis of no intervention effect (i.e., that it does not matter which is the actual point of change in phase). Checking the results for different possible definitions of an "immediate effect" is reasonable in case there is no clear expectation or justification regarding the number of measurements to use. It should be noted that we are *not* suggesting here to try several possible operational definitions and to report selectively only the most favorable one (see more in the next section). We do consider that a sensitivity analysis approach may be preferable to picking an operational definition without a clear (empirical or theoretical) basis and just hoping that the conclusions would be correct.

In summary, researchers need to be aware that their conclusions can be sensitive to the analytical decisions made (e.g., the way in which immediacy is operationally defined). In the absence of a gold standard, such analytical decisions need to be clearly justified (Tincani & Travers, 2022). If a single option cannot be reasonably justified, then the sensitivity analysis approach (Steegen et al., 2016) becomes potentially useful.

Understanding of the numerical results can be enhanced by visually inspecting the time series plot with the raw data (Kratochwill et al., 2021; Parker et al., 2006). Such a visual inspection can help identifying the onset and offset of the treatment effect, as well as the degree to which any unwanted or unexpected variability or trends can affect the quantifications. Finally, it should be noted that anything computed needs to be reported, as we further stress in the following section.

Sensitivity Analysis Approach and Reporting

Immediacy is not the only data aspect that is susceptible to multiple operational definitions. As mentioned in the introduction, the same is the case for overlap (Parker, Vannest, & Davis, 2011), level (e.g., choosing between a standardized mean or a percentage change measure), and trend (Manolov, 2018). Therefore, it is possible to perform a sensitivity analysis for each of these data aspects, if they are chosen to be the focus of the analysis, according to the type of effect expected. In case a single operational definition is chosen for the focal data feature, a justification is required. In that sense, we echo the recommendation from the Risk of Bias in N-of-1 Trials methodological quality scale (Tate et al., 2013) that in the absence of a clear consensus about the most appropriate data analytical approach, it is indispensable to justify the reason for choosing one option over the other.

A second aspect related to reporting refers to the need to report the results of all planned analyses (including test statistics in a randomization test that were chosen prior to gathering the data). This would enable avoiding selective reporting (Kratochwill et al., 2018; Tincani & Travers, 2018), which is a kind of questionable research practice that can take place in the SCED context (Laraway et al., 2019). Relatedly, it is worth discussing how to proceed when the initial expectations are not matched by the observed data pattern (including the results of the SART approach). To begin with, in case the researchers do not have a solid basis for stating a priori whether the effect of the intervention should be immediate or not, they could be explicit about the exploratory character of their analysis (including an exploratory application of the SART approach). In

[1] In this case it could be understood as identifying an operational definition of immediacy that is only appropriate (or only leads to large effects) for the data at hand.

case there are a priori expectations and these are not met by the observed data pattern (e.g., the effect is expected to be immediate but the results suggest it is delayed, or vice versa), there are four suggested actions to be taken: (a) be explicit, in the written report, about the divergence between the initial expectation and the data pattern observed; (b) explore whether the kind of effect observed is consistent across all replications within the study (e.g., all A-B comparisons that are part of a multiple-baseline or a reversal design); (c) review the available evidence from previous research regarding whether immediate effects have been previously observed; and (d) frame the interpretation of the results (in terms of the strength of the evidence for a causal relation between the intervention and the changes in the target behavior) according to the two previously mentioned aspects: the consistency of the results within the study and the degree of convergence with previous studies.

Finally, it should be noted that following the SART approach when assessing immediacy entails checking qualitatively the congruency of the results, and it is not equivalent to the multiple randomization testing procedure presented by Tanious, De, and Onghena (2019). These authors proposed testing for several effects simultaneously, while keeping the Type I familywise error rate under control.

Limitations

Despite the illustrations using mean difference, it should be noted that the SART approach is not restricted to focusing on level. We reiterate the importance of choosing the kind of quantification according to the expected effect (or data pattern), as commonly recommended in the context of randomization tests (Edgington, 1975; Levin et al., 2021). In the current text we did not provide details on how to select the appropriate quantification, but we here direct the interested reader to several potentially useful sources. For instance, if a regression (including multilevel) model is used, Moeyaert, Ugille, et al. (2014) and Natesan Bateley and Hedges (2021) provide indications on model building. For a nonparametric measure such as Tau (Parker, Vannest, Davis, & Sauber, 2011, see also Tarlow, 2017, for an alternative), Fingerhut et al. (2021) provide indications about how to choose one of the possible versions. When a randomization test is used, given that the test statistic can be chosen by the researchers, the discussion by Levin et al. (2021) is useful. Finally, covering several possible quantifications, including the ones previously mentioned in the current paragraph, Manolov et al. (2022) offer suggestions regarding the choice of a quantification according to the design and the expected data pattern.

In relation to any potential criticism directed towards *p*-values (e.g., Branch, 2014; J. Cohen, 1994), it should be noted that a *p*-value could be understood simply as the degree to which the observed difference is extreme, considering all possible differences (all potential moments of change in phase) and is just another way of expressing the rank that would be assigned to the actually observed difference. Thus, the focus would be placed on the rank of the quantification of immediate effect obtained for the actual moment of change in phase: The evidence for an immediate effect will be stronger whenever the rank is close to 1, regardless of the number of potential changes in phase. Therefore, the main use of the SART approach would be to compare whether the initial expectations (regarding whether the effect should be immediate) are matched by the results (regarding the moment of largest difference): Such a use is not bound to statistical power, as the aim is not necessarily to achieve $p \leq .05$. However, having few measurements entails having a smaller sample of the target behavior and few potential moments of change in phase, which may compromise obtaining clear distinctions between immediate and delayed effects, and changes taking place before the introduction of the intervention.

Future Research

In future research, we consider that a simulation study could be useful to inform about the statistical power of a randomization test using (a) different number of measurements when computing the test statistic; and (b), different test statistics (i.e., with different focal data features). Furthermore, a field test in a research domain where immediate effects are expected could be performed to obtain evidence regarding whether the conclusions of studies usually vary according to (a) the number of measurements used when computing immediate effects; and (b) the focal data feature.

Another line of research could focus on the way in which the SART approach can be integrated with (and contribute to) meta-analysis. As ad hoc suggestions, we can point at the possibility to count, for each study, whether the a priori expectations are well-aligned with the evidence from applying the SART approach. This could lead to a simple form of quantitative integration known as "vote-counting". Other options, related to randomization tests (although not specific to the SART approach), include the quantitative integration of the values of the test statistic (which can be effect size indices) or the combination of probabilities (Onghena et al., 2018).

References

Allison, D. B., & Gorman, B. S. (1993). Calculating effect sizes for meta-analysis: The case of the single case. *Behaviour Research and Therapy*, *31*(6), 621–631. https://doi.org/10.1016/0005-7967(93)90115-B

Aydin, O., & Tanious, R. (2022). Performance criteria-based effect size (PCES) measurement of single-case experimental designs: A real-world data study. *Journal of Applied Behavior Analysis*, *55*(3), 891−918. https://doi.org/10.1002/jaba.928

Baek, E. K., & Ferron, J. M. (2013). Multilevel models for multiple-baseline data: Modeling across-participant variation in autocorrelation and residual variance. *Behavior Research Methods*, *45*(1), 65–74. https://doi.org/10.3758/s13428-012-0231-z

Barber, A. B., Saffo, R. W., Gilpin, A. T., Craft, L. D., & Goldstein, H. (2016). Peers as clinicians: Examining the impact of stay play talk on social communication in young preschoolers with autism. *Journal of Communication Disorders*, *59*, 1–15. https://doi.org/10.1016/j.jcomdis.2015.06.009

Barker, J., McCarthy, P., Jones, M., & Moran, A. (2011). *Single case research methods in sport and exercise psychology*. Routlege.

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

Barnard-Brak, L., Richman, D. M., & Watkins, L. (2020). Treatment burst data points and single case design studies: A Bayesian N-of-1 analysis for estimating treatment effect size. *Perspectives on Behavior Science*, *43*(2), 285–301. https://doi.org/10.1007/s40614-020-00258-8

Barnard-Brak, L., Watkins, L., & Richman, D. (2021). Optimal number of baseline sessions before changing phases within single-case experimental designs. *Behavioural Processes*, *191*, 104461. https://doi.org/10.1016/j.beproc.2021.104461

Barton, E. E., Lloyd, B. P., Apriggs, A. D., & Gast, D. L. (2018). Visual analysis of graphic data. In J. R. Ledford & D. L. Gast (Eds.), *Single case research methodology: Applications in special education and behavioral sciences* (3rd ed.) (pp. 179–214). Routledge.

Barton, E. E., Meadan, H., & Fettig, A. (2019). Comparison of visual analysis, non-overlap methods, and effect sizes in the evaluation of parent implemented functional assessment based interventions. *Research in Developmental Disabilities*, *85*, 31–41. https://doi.org/10.1016/j.ridd.2018.11.001

Barton, E. E., Pustejovsky, J. E., Maggin, D. M., & Reichow, B. (2017). Technology-aided instruction and intervention for students with ASD: A meta-analysis using novel methods of estimating effect sizes for single-case research. *Remedial and Special Education*, *38*(6), 371–386. https://doi.org/10.1177/0741932517729508

Beretvas, S. N., & Chung, H. (2008). A review of meta-analyses of single-subject experimental designs: Methodological issues and practice. *Evidence-Based Communication Assessment and Intervention*, *2*(3), 129–141. https://doi.org/10.1080/17489530802446302

Branch, M. (2014). Malignant side effects of null-hypothesis significance testing. *Theory & Psychology*, *24*(2), 256–277. https://doi.org/10.1177/0959354314525282

Brogan, K. M., Rapp, J. T., & Sturdivant, B. R. (2019). Transition states in single case experimental designs. *Behavior Modification*. Advance online publication. https://doi.org/10.1177/0145445519839213

Brossart, D. F., Vannest, K., Davis, J., & Patience, M. (2014). Incorporating nonoverlap indices with visual analysis for quantifying intervention effectiveness in single-case experimental designs. *Neuropsychological Rehabilitation*, *24*(3-4), 464−491. https://doi.org/10.1080/09602011.2013.868361

Busk, P. L., & Serlin, R. C. (1992). Meta-analysis for single-case research. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research designs and analysis: New directions for psychology and education* (pp. 187−212). Lawrence Erlbaum.

Callahan, C. D., & Barisa, M. T. (2005). Statistical process control and rehabilitation outcome: The single-subject design reconsidered. *Rehabilitation Psychology*, *50*(1), 24–33. https://doi.org/10.1037/0090-5550.50.1.24

Carr, J. E. (2005). Recommendations for reporting multiple-baseline designs across participants. *Behavioral Interventions*, *20*(3), 219–224. https://doi.org/10.1002/bin.191

Center, B. A., Skiba, R. J., & Casey, A. (1985). A methodology for the quantitative synthesis of intra-subject design research. *The Journal of Special Education*, *19*(4), 387–400. https://doi.org/10.1177/002246698501900404

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

Chen, L.-T., Wu, P. J., & Peng, C.-Y. J. (2019). Accounting for baseline trends in intervention studies: Methods, effect sizes, and software. *Cogent Psychology*, *6*(1), 1679941. https://doi.org/10.1080/23311908.2019.1679941

Chen, M., Hyppa-Martin, J. K., Reichle, J. E., & Symons, F. J. (2016). Comparing single case design overlap-based effect size metrics from studies examining speech generating device interventions. *American Journal on Intellectual and Developmental Disabilities*, *121*(3), 169–193. https://doi.org/10.1352/1944-7558-121.3.169

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*(12), 997–1003. https://doi.org/10.1037/0003-066X.49.12.997

Cohen, L. L., Feinstein, A., Masuda, A., & Vowles, K. E. (2014). Single-case research design in pediatric

psychology: Considerations regarding data analysis. *Journal of Pediatric Psychology*, *39*(2), 124–137. https://doi.org/10.1093/jpepsy/jst065

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

Coulter, G. A., & Lambert, M. C. (2015). Access to general education curriculum: The effect of preteaching key words upon fluency and accuracy in expository text. *Learning Disabilities Quarterly*, *38*(4), 248–256. https://doi.org/10.1177/0731948715580438

Craig, A. R., & Fisher, W. W. (2019). Randomization tests as alternative analysis methods for behavior-analytic data. *Journal of the Experimental Analysis of Behavior*, *111*(2), 309–328. https://doi.org/10.1002/jeab.500

Dallery, J., & Raiff, B. R. (2014). Optimizing behavioral health interventions with single-case designs: From development to dissemination. *Translational Behavioral Medicine*, *4*(3), 290–303. https://doi.org/10.1007/s13142-014-0258-z

Edgington, E. S. (1967). Statistical inference from N=1 experiments. *The Journal of Psychology*, *65*(2), 195–199. https://doi.org/10.1080/00223980.1967.10544864

Edgington, E. S. (1975). Randomization tests for one-subject operant experiments. *The Journal of Psychology*, *90*(1), 57–68. https://doi.org/10.1080/00223980.1975.9923926

Edgington, E. S. (1996). Randomized single-subject experimental designs. *Behaviour Research and Therapy*, *34*(7), 567–574. https://doi.org/10.1016/0005-7967(96)00012-5

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

Epstein, L. H., Bickel, W. K., Czajkowski, S. M., Paluch, R. A., Moeyaert, M., & Davidson, K. W. (2021). Single case designs for early phase behavioral translational research in health psychology. *Health Psychology*, *40*(12), 858–874. https://doi.org/10.1037/hea0001055

Ferron, J. M., Moeyaert, M., Van den Noortgate, W., & Beretvas, S. N. (2014). Estimating causal effects from multiple-baseline studies: Implications for design and analysis. *Psychological Methods*, *19*(4), 493–510. https://doi.org/10.1037/a0037038

Fingerhut, J., Xu, X., & Moeyaert, M. (2021). Selecting the proper Tau-U measure for single-case experimental designs: Development and application of a decision flowchart. *Evidence-Based Communication Assessment and Intervention*, *15*(3), 99–114. https://doi.org/10.1080/17489539.2021.1937851

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

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

Franklin, R. D., Gorman, B. S., Beasley, T. M., & Allison, D. B. (1996). Graphical display and visual analysis. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 119–158). Lawrence Erlbaum.

Gage, N. A., & Lewis, T. J. (2013). Analysis of effect for single-case design research. *Journal of Applied Sport Psychology*, *25*(1), 46–60. https://doi.org/10.1080/10413200.2012.660673

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

Giannakakos, A. R., & Lanovaz, M. J. (2019). Using AB designs with nonoverlap effect size measures to support clinical decision-making: A Monte Carlo validation. *Behavior Modification*. Advance online publication. https://doi.org/10.1177/0145445519860219

Graham, J. E., Karmarkar, A. M., & Ottenbacher, K. J. (2012). Small sample research designs for evidence-based rehabilitation: Issues and methods. *Archives of Physical Medicine and Rehabilitation*, *93*(8), S111–S116. https://doi.org/10.1016/j.apmr.2011.12.017

Haegele, J. A., & Hodge, S. R. (2015). The applied behavior analysis research paradigm and single-subject designs in adapted physical activity research. *Adapted Physical Activity Quarterly*, *32*(4), 285–301. https://doi.org/10.1123/APAQ.2014-0211

Hamed, K. H., & Rao, A. A. (1998). A modified Mann-Kendall trend test for autocorrelated data. *Journal of Hydrology*, *204*(1–4), 182–196. https://doi.org/10.1016/S0022-1694(97)00125-X

Hedges, L. V., Pustejovsky, J. E., & Shadish, W. R. (2012). A standardized mean difference effect size for single case designs. *Research Synthesis Methods*, *3*(3), 224−239. https://doi.org/10.1002/jrsm.1052

Hedges, L. V., Pustejovsky, J. E., & Shadish, W. R. (2013). A standardized mean difference effect size for multiple baseline designs across individuals. *Research Synthesis Methods*, *4*(4), 324−341. https://doi.org/10.1002/jrsm.1086

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. (2014a). Analysis of single-case data: Randomisation tests for measures of effect size. *Neuropsychological Rehabilitation*, *24*(3-4), 507–527. https://doi.org/10.1080/09602011.2013.818564

Heyvaert, M., & Onghena, P. (2014b). Randomization tests for single-case experiments: State of the art, state of the science, and state of the application. *Journal of Contextual Behavioral Science*, *3*(1), 51–64. https://doi.org/10.1016/j.jcbs.2013.10.002

Hitchcock, J. H., Kratochwill, T. R., & Chezan, L. C. (2015). What Works Clearinghouse standards and generalization of single-case design evidence. *Journal of Behavioral Education*, *24*(4), 459-469. https://doi.org/10.1007/s10864-015-9224-1

Holcombe, A., Wolery, M., & Gast, D. L. (1994). Comparative single subject research: Description of designs and discussion of problems. *Topics in Early Childhood and Special Education*, *16*(1), 168–190. https://doi.org/10.1177/027112149401400111

Holman, G., & Koerner, K. (2014). Single case designs in clinical practice: A contemporary CBS perspective on

why and how to. *Journal of Contextual Behavioral Science*, *3*(2), 138–147. https://doi.org/10.1016/j.jcbs.2014.04.006

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

Horner, R. H., & Kratochwill, T. R. (2012). Synthesizing single-case research to identify evidence-based practices: Some brief reflections. *Journal of Behavioral Education*, *21*(3), 266–272. https://doi.org/10.1007/s10864-012-9152-2

Horner, R. J., & Odom, S. L. (2014). Constructing single-case research designs: Logic and options. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: Methodological and statistical advances* (pp. 27–51). American Psychological Association. https://doi.org/10.1037/14376-002

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

Huitema, B. E., & McKean, J. W. (1991). Autocorrelation estimation and inference with small samples. *Psychological Bulletin*, *110*(2), 291–304. https://doi.org/10.1037/0033-2909.110.2.291

Jacobs, K. W. (2019). Replicability and randomization test logic in behavior analysis. *Journal of the Experimental Analysis of Behavior*, *111*(2), 329–341. https://doi.org/10.1002/jeab.501

Jamshidi, L., Heyvaert, M., Declercq, L., Fernández-Castilla, B., Ferron, J. M., Moeyaert, M., Beretvas, S. N., Onghena, P., & Van den Noortgate, W. (2022). A systematic review of single-case experimental design meta-analyses: Characteristics of study designs, data, and analyses. *Evidence-Based Communication Assessment and Intervention*. Advance online publication. https://doi.org/10.1080/17489539.2022.2089334

Janosky, J. E., Leininger, S. L., Hoerger, M. P., & Libkuman, T. M. (2009). *Single subject designs in biomedicine*. Springer.

Katz, B. R., & Lattal, K. A. (2021). What is an extinction burst?: A case study in the analysis of transitional behavior. *Journal of the Experimental Analysis of Behavior*, *115*(1), 129–140. https://doi.org/10.1002/jeab.642

Kazdin, A. E. (2020). *Single-case research designs: Methods for clinical and applied settings* (3rd ed.). Oxford University Press.

Kazdin, A. E. (2021). Single-case experimental designs: Characteristics, changes, and challenges. *Journal of the Experimental Analysis of Behavior*, *115*(1), 56–85. https://doi.org/10.1002/jeab.638

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

Kilgus, S. P., Riley-Tillman, T. C., & Kratochwill, T. R. (2016). Establishing interventions via a theory-driven single case design research cycle. *School Psychology Review*, *45*(4), 477–498. https://doi.org/10.17105/SPR45-4.477-498

Kipfmiller, K. J., Brodhead, M. T., Wolfe, K., LaLonde, K., Sipila, E. S., Bak, M. S., & Fisher, M. H. (2019). Training front-line employees to conduct visual analysis using a clinical decision-making model. *Journal of Behavioral Education*, *28*(3), 301–322. https://doi.org/10.1007/s10864-018-09318-1

Krasny-Pacini, A., & Evans, J. (2018). Single-case experimental designs to assess intervention effectiveness in rehabilitation: A practical guide. *Annals of Physical and Rehabilitation Medicine*, *61*(3), 164–179. https://doi.org/10.1016/j.rehab.2017.12.002

Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2013). Single-case intervention research design standards. *Remedial and Special Education*, *34*(1), 26–38. https://doi.org/10.1177/0741932512452794

Kratochwill, T. R., Horner, R. H., Levin, J. R., Machalicek, W., Ferron, J., & Johnson, A. (2021). Single-case design standards: An update and proposed upgrades. *Journal of School Psychology*, *89*, 91–105. https://doi.org/10.1016/j.jsp.2021.10.006

Kratochwill, T. R., & Levin, J. R. (2010). Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. *Psychological Methods*, *15*(2), 124–144. https://doi.org/10.1037/a0017736

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

Kratochwill, T. R., Levin, J. R., & Horner, R. H. (2018). Negative results: Conceptual and methodological dimensions in single-case intervention research. *Remedial and Special Education*, *34*(1), 26–38. https://doi.org/10.1177/0741932512452794

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

Krone, T., Albers, C. J., & Timmerman, M. E. (2016). Comparison of estimation procedures for multilevel AR (1) models. *Frontiers in Psychology*, *7*, 486. https://doi.org/10.3389/fpsyg.2016.00486

Kwasnicka, D., Inauen, J., Nieuwenboom, W., Nurmi, J., Schneider, A., Short, C. E., Dekkers, T., Williams, A. J., Bierbauer, W., Haukkala, A., Picariello, F., & Naughton, F. (2019). Challenges and solutions for N-of-1 design studies in health psychology. *Health Psychology Review*, *13*(2), 163–178. https://doi.org/10.1080/17437199.2018.1564627

Lambert, M. C., Cartledge, G., Heward, W. L., & Lo, Y.-Y. (2006). Effects of response cards on disruptive behavior and academic responding during math lessons by fourth-grade urban students. *Journal of Positive Behavior Interventions*, *8*(2), 88–99. https://doi.org/10.1177/10983007060080020701

Lancioni, G. E., Desideri, L., Singh, N. N., Sigafoos, J., & O'Reilly, M. F. (2021). A commentary on standards for single-case experimental studies. *International Journal of Developmental Disabilities*. Advance online publication. https://doi.org/10.1080/20473869.2020.1870420

Lane, J. D., & Gast, D. L. (2014). Visual analysis in single case experimental design studies: Brief review and guidelines. *Neuropsychological Rehabilitation*, *24*(3-4), 445–463. https://doi.org/10.1080/09602011.2013.815636

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

Lane, J. D., Shepley, C., & Spriggs, A. D. (2021). Issues and improvements in the visual analysis of A-B single-case graphs by pre-service professionals. *Remedial and Special Education*, *42*(4), 235–247. https://doi.org/10.1177/0741932519873120

Lanovaz, M. J., & Turgeon, S. (2020). How many tiers do we need? Type I errors and power in multiple baseline designs. *Perspectives on Behavior Science*, *43*(3), 605–616. https://doi.org/10.1007/s40614-020-00263-x

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

Laraway, S., Snycerski, S., Pradhan, S., & Huitema, B. E. (2019). An overview of scientific reproducibility: Consideration of relevant issues for behavior science/analysis. *Perspectives on Behavior Science*, *42*(1), 33–57. https://doi.org/10.1007/s40614-019-00193-3

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. (2022). Concurrence on nonconcurrence in multiple-baseline designs: A commentary on Slocum et al. (2022). *Perspectives on Behavior Science*. Advance online publication. https://doi.org/10.1007/s40614-022-00342-1

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. (Eds.) (2018). *Single case research methodology: Applications in special education and behavioral sciences* (3rd ed.) Routledge.

Ledford, J. R., Lane, J. D., & Severini, K. E. (2018). Systematic use of visual analysis for assessing outcomes in single case design studies. *Brain Impairment*, *19*(1), 4–17. https://doi.org/10.1017/BrImp.2017.16

Lenz, A. S. (2013). Calculating effect size in single-case research: A comparison of nonoverlap methods. *Measurement and Evaluation in Counseling and Development*, *46*(1), 64–73. https://doi.org/10.1177/0748175612456401

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

Levin, J. R., Ferron, J. M., & Gafurov, B. S. (2021). Investigation of single-case multiple-baseline randomization tests of trend and variability. *Educational Psychology Review*, *33*(2), 713–737. https://doi.org/10.1007/s10648-020-09549-7

Lieberman, R. G., Yoder, P. J., Reichow, B., & Wolery, M. (2010). Visual analysis of multiple baseline across participants graphs when change is delayed. *School Psychology Quarterly*, *25*(1), 28–44. https://doi.org/10.1037/a0018600

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

Macpherson, K., Charlop, M. H., & Miltenberger, C. A. (2015). Using portable video modeling technology to increase the compliment behaviors of children with autism during athletic group play. *Journal of Autism and Developmental Disorders*, *45*(12), 3836–3845. https://doi.org/10.1007/s10803-014-2072-3

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., 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

Maggin, D. M., O'Keefe, B. V., & Johnson, A. H. (2011). A quantitative synthesis of methodology in the meta-analysis of single-subject research for students with disabilities: 1985–2009. *Exceptionality*, *19*(2), 109–135. https://doi.org/10.1080/09362835.2011.565725

Maggin, D. M., Pustejovsky, J. E., & Johnson, A. H. (2017). A meta-analysis of school-based group contingency interventions for students with challenging behavior: An update. *Remedial and Special Education*, *38*(6), 353–370. https://doi.org/10.1177/0741932517717043

Maggin, D. M., Swaminathan, H., Rogers, H. J., O'Keefe, B. V., Sugai, G., & Horner, R. H. (2011). A generalized least squares regression approach for computing effect sizes in single-case research Application examples. *Journal of School Psychology*, *49*(3), 301–321. https://doi.org/10.1016/j.jsp.2011.03.004

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., & Ferron, J. M. (2020). Assessing consistency of effects when applying multilevel models to single-case data. *Behavior Research Methods*, *52*(6), 2460–2479. https://doi.org/10.3758/s13428-020-01417-0

Manolov, R., Moeyaert, M., & Fingerhut, J. (2022). A priori justification for effect measures in single-case experimental designs. *Perspectives on Behavior Science*, *45*(1), 156–189. https://doi.org/10.1007/s40614-021-00282-2

Manolov, R., Solanas, A., & Sierra, V. (2019). Extrapolating baseline trend in single-case data: Problems and tentative solutions. *Behavior Research Methods*, *51*(6), 2847–2869. https://doi.org/10.3758/s13428-018-1165-x

Manolov, R., Solanas, A., & Sierra, V. (2020). Changing criterion designs: Integrating methodological and data analysis recommendations. *The Journal of Experimental Education*, *88*(2), 335-350. https://doi.org/10.1080/00220973.2018.1553838

Manolov, R., & Tanious, R. (2022). Assessing consistency in single-case data features using modified Brinley

plots. *Behavior Modification*, *46*(3), 581–627. https://doi.org/10.1177/0145445520982969

Manolov, R., Tanious, R., De, T. K., & Onghena, P. (2021). Assessing consistency in single-case alternation designs. *Behavior Modification*, *45*(6), 929–961. https://doi.org/10.1177/0145445520923990

Manolov, R., & Vannest, K. (2019). A visual aid and objective rule encompassing the data features of visual analysis. *Behavior Modification*. Advance online publication. https://doi.org/10.1177/0145445519854323

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

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

Miller, M. J. (1985). Analyzing client change graphically. *Journal of Counseling and Development*, *63*(8), 491-494. https://doi.org/10.1002/j.1556-6676.1985.tb02743.x

Miočević, M., Klaassen, F., Geuke, G., Moeyaert, M., & Maric, M. (2020). Using Bayesian methods to test mediators of intervention outcomes in single case experimental designs. *Evidence-Based Communication Assessment and Intervention*, *14*(1-2), 52-68. https://doi.org/10.1080/17489539.2020.1732029

Moeyaert, M., Ferron, J., Beretvas, S., & Van den Noortgate, W. (2014). From a single-level analysis to a multilevel analysis of since-case experimental designs. *Journal of School Psychology*, *52*(2), 191−211. https://doi.org/10.1016/j.jsp.2013.11.003

Moeyaert, M., Ugille, M., Ferron, J., Beretvas, S. N., & Van den Noortgate, W. (2014). The influence of the design matrix on treatment effect estimates in the quantitative analyses of single-case experimental designs research. *Behavior Modification*, *38*(5), 665–704. https://doi.org/10.1177/0145445514535243

Morgan, D. L., & Morgan, R. K. (2009). *Single-case research methods for the behavioral and health sciences*. Sage.

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

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

Natesan, P., & Hedges, L. V. (2017). Bayesian unknown change-point models to investigate immediacy in single case designs. *Psychological Methods*, *22*(4), 743–759. https://doi.org/10.1037/met0000134

Natesan Batley, P., Contractor, A. A., & Caldas, S. V. (2020). Bayesian time-series models in single case experimental designs: A tutorial for trauma researchers. *Journal of Traumatic Stress*, *33*(6), 1144–1153. https://doi.org/10.1002/jts.22614

Natesan Batley, P., & Hedges, L. V. (2021). Accurate models vs. accurate estimates: A simulation study of Bayesian single-case experimental designs. *Behavior Research Methods*, *53*(4), 1782–1798. https://doi.org/10.3758/s13428-020-01522-0

Natesan Batley, P., Minka, T. & Hedges, L. V. (2020). Investigating immediacy in multiple-phase-change single-case experimental designs using a Bayesian unknown change-points model. *Behavior Research Methods*, *52*(4), 1714–1728. https://doi.org/10.3758/s13428-020-01345-z

Natesan Batley, P., Nandakumar, R., Palka, J. M., & Shrestha, P. (2021). Comparing the Bayesian unknown change-point model and simulation modeling analysis to analyze single case experimental designs. *Frontiers in Psychology*, *11*, article 617047. https://doi.org/10.3389/fpsyg.2020.617047

Ninci, J. (2019). Single-case data analysis: A practitioner guide for accurate and reliable decisions. *Behavior Modification*. Advance online publication. https://doi.org/10.1177/0145445519867054

Odom, S. L., Barton, E. E., Reichow, B., Swaminathan, H., & Pustejovsky, J. E. (2018). Between-case standardized effect size analysis of single case designs: Examination of the two methods. *Research in Developmental Disabilities*, *79*(1), 88–96. https://doi.org/10.1016/j.ridd.2018.05.009

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. (1992). Randomization tests for extensions and variations of ABAB single-case experimental designs: A rejoinder. *Behavioral Assessment*, *14*(2), 153–171.

Onghena, P., Michiels, B., Jamshidi, L., Moeyaert, M., & Van den Noortgate, W. (2018). One by one: Accumulating evidence by using meta-analytical procedures for single-case experiments. *Brain Impairment*, *19*(1), 33–58. https://doi.org/10.1017/BrImp.2017.25

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

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., & Davis, J. L. (2011). Effect size in single-case research: A review of nine nonoverlap techniques. *Behavior Modification*, *35*(4), 303–322. https://doi.org/10.1177/0145445511399147

Parker, R. I., Vannest, K. J., & Davis, J. L. (2014). A simple method to control positive baseline trend within data nonoverlap. *The Journal of Special Education*, *48*(2), 79–91. https://doi.org/10.1177/0022466912456430

Parker, R. I., Vannest, K. J., Davis, J. L., & Sauber, S. B. (2011). Combining nonoverlap and trend for single-case research: Tau-U. *Behavior Therapy*, *42*(2), 284−299. https://doi.org/10.1016/j.beth.2010.08.006

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

Parsonson, B. S., & Baer, D. M. (1986). The graphic analysis of data. In A. Poling & R. W. Fuqua (Eds.), *Research methods in applied behavior analysis: Issues and advances* (pp. 157–186). Plenum Press.

Peng, C. Y. J., & Chen, L.-T. (2018). Handling missing data in single-case studies. *Journal of Modern Applied Statistical Methods*, *17*(1), eP2488. https://doi.org/10.22237/jmasm/1525133280

Petursdottir, A. I., & Carr, J. E. (2018). Applying the taxonomy of validity threats from mainstream research design to single-case experiments in applied behavior

analysis. *Behavior Analysis in Practice*, *11*(3), 228-240. https://doi.org/10.1007/s40617-018-00294-6

Pfadt, A., & Wheeler, D. J. (1995). Using statistical process control to make data-based clinical decisions. *Journal of Applied Behavior Analysis*, *28*(3), 349–370. https://doi.org/10.1901/jaba.1995.28-349

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

Poling, A., & Fuqua, R. W. (Eds.) (1986). *Research methods in applied behavior analysis: Issues and advances*. Plenum Press.

Pustejovsky, J. E. (2018). Using response ratios for meta-analyzing single-case designs with behavioral outcomes. *Journal of School Psychology*, *68*, 99−112. https://doi.org/10.1016/j.jsp.2018.02.003

Pustejovsky, J. E., Swan, D. M., & English, K. W. (2019). An examination of measurement procedures and characteristics of baseline outcome data in single-case research. *Behavior Modification*. Advance online publication. https://doi.org/10.1177/0145445519864264

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

Rakap, S. (2015). Effect sizes as result interpretation aids in single-subject experimental research: Description and application of four nonoverlap methods. *British Journal of Special Education*, *42*(1), 11–33. https://doi.org/10.1111/1467-8578.12091

Riley-Tillman, T. C., Burns, M. K., & Kilgus, S. P. (2020). *Evaluating educational interventions: Single-case design for measuring response to intervention* (2nd ed.). The Guilford Press.

Rindskopf, D. (2014). Bayesian analysis of data from single case designs. *Neuropsychological Rehabilitation*, 24(3-4), 572–589. https://doi.org/10.1080/09602011.2013.866903

Scruggs, T. E., Mastropieri, M. A., & Casto, G. (1987). The quantitative synthesis of single-subject research: Methodology and validation. *Remedial and Special Education*, *8*(2), 24–33. https://doi.org/10.1177/074193258700800206

Shadish, W. R. (2014). Analysis and meta-analysis of single-case designs: An introduction. *Journal of School Psychology*, *52*(2), 109–122. https://doi.org/10.1016/j.jsp.2013.11.009

Shadish, W. R., Hedges, L. V., Horner, R. H., & Odom, S. L. (2015). *The role of between-case effect size in conducting, interpreting, and summarizing single-case research* (NCER-2015-02). National Center for Education Research, Institute of Education Sciences, U.S. Department of Education. Retrieved from http://ies.ed.gov/ncser/pubs/2015002/.

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., Rindskopf, D. M., Hedges, L. V., & Sullivan, K. J. (2013). Bayesian estimates of autocorrelations in single-case designs. *Behavior Research Methods*, *45*(3), 813–821. https://doi.org/10.3758/s13428-012-0282-1

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

Sidman, M. (1960). *Tactics of scientific research*. Basic Books.

Slocum, T. A., Pinkelman, S. E., Joslyn, P. R., & Nichols, B. (2022). Threats to internal validity in multiple-baseline design variations. *Perspectives on Behavior Science*. Advance online publication. https://doi.org/10.1007/s40614-022-00326-1

Smith, J. D. (2012). Single-case experimental designs: A systematic review of published research and current standards. *Psychological Methods*, *17*(4), 510–550. https://doi.org/10.1037/a0029312

Spear, C. F., Strickland-Cohen, M. K., Romer, N., & Albin, R. W. (2013). An examination of social validity within single-case research with students with emotional and behavioral disorders. *Remedial and Special Education*, *34*(6), 357–370. https://doi.org/10.1177/0741932513490809

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702–712. https://doi.org/10.1177/1745691616658637

Sullivan, K. J., Shadish, W. R., & Steiner, P. M. (2015). An introduction to modeling longitudinal data with generalized additive models: Applications to single-case designs. *Psychological Methods*, *20*(1), 26–42. https://doi.org/10.1037/met0000020

Swaminathan, H., Rogers, H. J., Horner, R., Sugai, G., & Smolkowski, K. (2014). Regression models for the analysis of single case designs. *Neuropsychological Rehabilitation*, *24*(3-4), 554−571. https://doi.org/10.1080/09602011.2014.887586

Swan, D. M., & Pustejovsky, J. E. (2018). A gradual effects model for single-case designs. *Multivariate Behavioral Research*, *53*(4), 574–593. https://doi.org/10.1080/00273171.2018.1466681

Swan, D. M., Pustejovsky, J. E., & Beretvas, S. N. (2020). The impact of response-guided designs on count outcomes in single-case experimental design baselines. *Evidence-Based Communication Assessment and Intervention*, *14*(1-2), 82–107. https://doi.org/10.1080/17489539.2020.1739048

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

Tanious, R., Manolov, R., & Onghena, P. (2021). The assessment of consistency in single-case experiments: Beyond A-B-A-B designs. *Behavior Modification*, *45*(4), 560–580. https://doi.org/10.1177/0145445519882889

Tanious, R., & Onghena, P. (2021). A systematic review of applied single-case research published between 2016 and 2018: Study designs, randomization, data aspects, and data analysis. *Behavior Research Methods*, *53*(4), 1371-1384. https://doi.org/10.3758/s13428-020-01502-4

Tankersley, M., McGoey, K. E., Dalton, D., Rumrill Jr, P. D., & Balan, C. M. (2006). Single subject research methods in rehabilitation. *Work*, *26*(1), 85–92. PMID: 16373983

Tarlow, K. (2017). An improved rank correlation effect size statistic for single-case designs: Baseline corrected Tau. *Behavior Modification*, *41*(4), 427–467. https://doi.org/10.1177/0145445516676750

Tarlow, K. R., & Brossart, D. F. (2018). A comprehensive method of single-case data analysis: Interrupted Time-Series Simulation (ITSSIM). *School Psychology Quarterly*, *33*(4), 590-603. https://doi.org/10.1037/spq0000273

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., McDonald, S., Togher, L., & Rosenkoetter, U. (2014). The conduct and report of single-case research: Strategies to improve the quality of the neurorehabilitation literature. *Neuropsychological Rehabilitation*, *24*(3–4), 315–331. https://doi.org/10.1080/09602011.2013.875043

Tate, R. L., Perdices, M., Rosenkoetter, U., Wakim, D., Godbee, K., Togher, L., & McDonald, S. (2013). Revision of a method quality rating scale for single-case experimental designs and n-of-1 trials: The 15-item Risk of Bias in N-of-1 Trials (RoBiNT) Scale. *Neuropsychological Rehabilitation*, *23*(5), 619–638. https://doi.org/10.1080/09602011.2013.824383

Tincani, M., & Travers, J. (2018). Publishing single-case research design studies that do not demonstrate experimental control. *Remedial and Special Education*, *39*(2), 118–128. https://doi.org/10.1177/0741932517697447

Tincani, M., & Travers, J. C. (2022). Questionable research practices in single-case experimental designs: Examples and possible solutions. In W. O'Donohue, A. Masuda, & S. O. Lilienfeld (Eds.), *Avoiding questionable research practices in applied psychology* (pp. 269-285). Springer Publications. https://doi.org/10.1007/978-3-031-04968-2_12

van de Schoot, R., & Miočević, M. (Eds.). (2020). *Small sample size solutions: A guide for applied researchers and practitioners*. Routledge.

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

Verboon, P., & Peters, G. J. (2020). Applying the generalized logistic model in single case designs: Modeling treatment-induced shifts. *Behavior Modification*, *44*(1), 27–48. https://doi.org/10.1177/0145445518791255

What Works Clearinghouse (2017). What Works Clearinghouse Standards Handbook, Version 4.0. U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Retrieved from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_standards_handbook_v4.pdf

What Works Clearinghouse (2020). What Works Clearinghouse Standards Handbook, Version 4.1. U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. Retrieved from https://ies.ed.gov/ncee/wwc/handbooks

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., van Aert, R. C., & Van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid *p*-hacking. *Frontiers in Psychology*, *7*, article 1832. https://doi.org/10.3389/fpsyg.2016.01832

Winkens, I., Ponds, R., Pouwels-van den Nieuwenhof, C., Eilander, H., & 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. *Neuropsychological Rehabilitation*, *24*(3–4), 349–364. https://doi.org/10.1080/09602011.2014.901229

Wolery, M. (2013). A commentary: Single-case design technical document of the What Works Clearinghouse. *Remedial and Special Education*, *34*(1), 39–43. https://doi.org/10.1177/0741932512468038

Wolery, M., Busick, M., Reichow, B., & Barton, E. E. (2010). Comparison of overlap methods for quantitatively synthesizing single-subject data. *The Journal of Special Education*, *44*(1), 18–29. https://doi.org/10.1177/0022466908328009

Wolfe, K., Barton, E. E., & Meadan, H. (2019). Systematic protocols for the visual analysis of single-case research data. *Behavior Analysis in Practice*, *12*(2), 491–502. https://doi.org/10.1007/s40617-019-00336-7

Received: March 16, 2022
Final Acceptance: August 28, 2022
Editor-in-Chief: Mark Galizio
Associate Editor: Brent Alsop